MEMORY AND PROBABILITY

Pedro Bordalo John J. Conlon Nicola Gennaioli
Spencer Y. Kwon Andrei Shleifer[1]

May 12, 2022

Abstract

In many economic decisions people estimate probabilities, such as the likelihood that a risk materializes or that a job applicant will be a productive employee, by retrieving experiences from memory. We model this process based on two established regularities of selective recall: similarity and interference. We show that the similarity structure of a hypothesis and the way it is described (not just its objective probability) shape the recall of experiences and thus probability assessments. The model accounts for and reconciles a variety of empirical findings, such as overestimation of unlikely events when these are cued versus neglect of non-cued ones, the availability heuristic, the representativeness heuristic, conjunction and disjunction fallacies, as well as over versus underreaction to information in different situations. The model yields several new predictions, for which we find strong experimental support.

[1] Saïd Business School, University of Oxford, Harvard University, Bocconi University and IGIER, Harvard University, and Harvard University. We are grateful to Ben Enke, Drew Fudenberg, Sam Gershman, Thomas Graeber, Cary Frydman, Lawrence Jin, Yueran Ma, Fabio Maccheroni, Sendhil Mullainathan, Salvo Nunnari, Dev Patel, Kunal Sangani, Jesse Shapiro, Josh Schwartzstein, Adi Sunderam, and Michael Woodford for helpful comments. Julien Manili provided outstanding research assistance. Gennaioli thanks the Italian Ministry of Education, University and Research, for financial support (PRIN 2017 Prot. 2017CY3SSY).

1. Introduction

It is well known that memory plays an important role in belief formation. Tversky and Kahneman (1973) show that when instances of a probabilistic hypothesis are easier to recall, the hypothesis is judged to be more likely, a finding they call the availability heuristic. When prompted to think about an unlikely event, such as dying in a tornado, people overestimate its frequency (Lichtenstein et al. 1978). They also attach a higher probability to an event if its description is broken down into constituent parts, which facilitates retrieval of instances (Fischoff et al. 1978). More broadly, beliefs depend on recalled personal experiences, such as stock market crashes (Malmendier and Nagel 2011), and not just on statistical information. Despite this evidence, a systematic analysis of the role of human memory in belief formation is lacking.

It is also well known that beliefs depart from rationality in a variety of ways, which shape behavior. Sometimes unlikely events are overestimated, as when consumers overpay for insurance (Sydnor 2010; Barseghyan et al. 2013) or bet in long-shot lotteries (Chiappori et al. 2019). Other times, unlikely events are underestimated, as when investors neglect tail risk (Gennaioli et al. 2012). Adding to the clutter, in finance there is abundant evidence of both over and underreaction to news. Beliefs overestimate the future earnings of individual firms and of the market after periods of rapid earnings growth (Bordalo et al. 2019, 2022), leading to long-run return reversals, but underestimate the impact of other news, such as earnings surprises, leading to return momentum (Chan et al. 1996; Bouchaud et al. 2019; Kwon and Tang 2021). This bewildering diversity of biases is puzzling, and has led sceptics to minimize the evidence on beliefs and stick to rationality.

In this paper, we show that building a theory a belief formation based on the psychology of human memory helps reconcile seemingly contradictory biases and generates new predictions. We test these predictions in two experiments and find substantial support.

In our theory, a decision maker (DM) estimates the frequency of a hypothesis H_1 relative to a disjoint alternative H_2 by the ease with which experiences of H_1 are retrieved from memory compared to experiences of H_2. In our stylized setting, the DM does not use any statistical information such as base rates. Instead, when the DM retrieves experiences of H_i, that hypothesis acts as a memory cue, shaping selective recall according to two well-established forces in the psychology of memory: similarity and interference (Kahana 2012).

To see these forces at work, consider a DM assessing the probability of H_1 = "cause of death is flood" compared to H_2 = "other causes of death". Similarity means that, when thinking about H_1, experiences of floods are easier to retrieve than those of earthquakes, because the former are more similar to the hypothesis than the latter. Interference means that, because we don't control what we recall, irrelevant experiences similar enough to the hypothesis or frequent enough in the database may be erroneously retrieved. When thinking about H_1 = "cause of death is flood", the DM may retrieve "accidental drowning", which belongs to the alternative H_2 = "other causes of death," or may retrieve "survival in a flood," which is a similar but irrelevant non-lethal experience.

When irrelevant experiences come to mind, they crowd out the retrieval of the relevant ones. This affects judgments: the hypothesis facing less interference is oversampled compared to its alternative, so its probability estimate is inflated. The key new implication is that whether a probability is over or under-estimated depends on the similarity structure of alternative hypotheses.

This mechanism reconciles conflicting findings and makes novel predictions. A cued unlikely event such as H_1 = "cause of death is flood" is overestimated for two reasons. First, cueing floods causes retrieval of similar events, leading to oversampling relative to their rare occurrence. Second, the alternative category H_2 = "causes of death other than flood" consists of many dissimilar events that are difficult to retrieve. As a result, broad and heterogeneous

hypotheses tend to be underestimated. But similarity also predicts that DM's should neglect rare events when these belong to a heterogeneous category that is not cued explicitly. When thinking about $H_2 =$ "causes of death other than flood" we may neglect airplane crashes because they are too dissimilar from typical causes of death, and so do not come to mind. Unlike Kahneman and Tversky's (1979) Prospect Theory, our model predicts when low probability events are over or underestimated, helping explain conflicting risk attitudes documented in the field.

Similarity also explains partition dependence: the probability estimate of the residual hypothesis "other causes of death" increases if it is instead formulated as H_2' = "cancer, heart attack, or other causes of death" (Tversky and Koehler 1994). Intuitively, the finer partition singles out subgroups whose elements are similar to each other but dissimilar to other experiences, which reduces interference and boosts recall.

In this way, similarity and interference account for many biases typically attributed to the availability heuristic. Strikingly, they also produce biases in conditional probability assessments that are typically attributed to representativeness (Kahneman and Tversky 1973). Consider the well-known base-rate neglect. When given the data that Steve is shy and withdrawn, subjects think he is more likely to be a librarian than a farmer, neglecting the fact that farmers are far more numerous than librarians. Similarity explains the mistake. It is easy to imagine Steve as H_1 = "a librarian," due to selective recall of many similarly shy librarians. It is in contrast harder to imagine shy farmers, because many farmers are not shy. Experiences of outgoing farmers, even though irrelevant for the assessment at hand, come to mind and interfere, reducing the belief that a shy person is a farmer. This same mechanism accounts for the conjunction fallacy (Tversky and Kahneman 1974, 1983). Similarity and interference help explain not only what people are likely to think about, but also what they systematically fail to think about.

These mechanisms shed new light on several economic applications. They explain both over and underreaction to data in conditional assessments, which have been documented in the lab and in the field, making predictions for when each arises. They help explain why people planning for retirement systematically fail to recall or imagine unlikely and heterogeneous spending shocks and hence under-save (Augenblick et al. 2022) and why people prefer to buy insurance against specific risks instead of the more convenient broad insurance (Kuhnreuther and Pauly 2006). Memory also offers a foundation for the kernel of truth model of social stereotypes (BCGS 2016), but yields a key new implication that stereotypes are especially inaccurate for minorities, because interference from the larger, majority group is stronger. In fact, minority stereotypes may be based on purely illusory correlations (Sherman, Hamilton, and Roskos-Ewoldsen 1989).

Finally, we test the new predictions of our model using a novel experimental design in which participants see 40 images that differ in content and in some cases also in color. Subjects then assess the probability that a randomly selected image possesses a certain property. To do so, they only need to recall what they saw. We manipulate the subjects' database of experiences and the cues they face when assessing a hypothesis. We also measure the recall of experiences. We find support for our predictions for how over and underestimation of unlikely events can be switched on and off by modulating similarity and interference. We also generate over and underreaction to data by varying the strength of the signal and the likelihood of the hypothesis. Across all treatments, recall of experiences and probability judgments are strongly correlated.

Recent research explores the role of memory in belief formation (Mullainathan 2002; Bordalo et al 2020; Wachter and Kahana 2019; Enke et al. 2020). Some see this phenomenon as efficient information processing (Tenenbaum and Griffiths 2001; Dasgupta et al. 2020; Azeredo da Silveira et al. 2020; Dasgupta and Gershman 2021). We instead start with well-documented

regularities in recall, and show how they unify the representativeness and availability heuristics (Tversky and Kahneman 1974). Due to similarity and interference, representative experiences are more "available," or accessible, for recall. This approach micro-founds and generalizes previous formalizations of representativeness (Gennaioli and Shleifer 2010; Bordalo et al. 2016), as well as the overreaction mechanism in Diagnostic Expectations (Bordalo et al. 2018).

Bordalo et al. (2020) present and experimentally test a model of memory-based beliefs in which the representativeness heuristic follows from a context dependent similarity function (Tversky 1977), meaning that the similarity of an experience to a hypothesis is higher if that experience is less likely in alternative hypotheses. Here we instead use a standard similarity function and obtain this effect as a special case of the broader role of interference in recall. This approach yields many new results. It yields biases attributed to the availability heuristic, and new predictions on the underestimation of heterogeneous hypotheses, the over vs underestimation of unlikely events, and the coexistence of under and over-reaction to data. We experimentally test these novel implications by extending the design in Bordalo et al. (2020).

In psychology, Sanborn and Chater (2016) present a model of beliefs based on Bayesian memory sampling. The Minerva-DM model (Dougherty et al. 1999) features similarity-based recall and noisy encoding, but does not allow for interference. By neglecting the joint action of similarity and interference, these models cannot account for key biases such as representativeness or the conjunction fallacy without making ad hoc ancillary assumptions. In Billot et al. (2005), the probability of an elementary event (rather than of broad hypotheses) is estimated based on its similarity to other events in the database. They do not study judgment biases and their model generates neither the conjunction nor the disjunction effect. In Johnson et al. (2007) query theory, buyers and sellers sample different aspects of a good from memory depending on how they are

cued with different queries. While they focus on explaining the endowment effect rather than probability biases, they also emphasize interference and similarity to the cue in shaping retrieval.

We describe our model of similarity-based recall and probability judgments in Section 2. We characterize the departures of probability estimates from statistically correct beliefs in Section 3. We present the experimental results in Sections 4. Section 5 discusses economic applications to savings and stereotypes. Section 6 concludes.

2. The Model

A Decision Maker's (DM) memory database E consists of $N > 1$ experiences, accumulated either through personal events, or via communication or media reports. An experience e is described by $F > 1$ features, each of which takes a value in $\{0,1\}$.

In our running example, we consider a database of potential causes of death. Here a subset of features captures different potential causes: f_1 may identify "car accident", f_2 "flood", f_3 "heart attack", etc. One feature, which we denote by f_d, indicates whether the event was lethal or not. There are superordinate features, such as f_{d+1} ="disease", f_{d+2} ="natural disaster", etc, which take the value of 1 for the relevant subsets of possible death events. Experiences are vectors of features. For instance, lethal heart attacks have $f_1 = f_2 = 0$, $f_3 = f_d = f_{d+1} = 1$ and $f_{d+2} = 0$. Non-lethal heart attacks have the same feature values except for $f_d = 0$. Additional features may include the characteristics of people involved, such as their age or gender, or contextual factors such as the time and emotion associated with the experience. The set of features is sufficiently large that no two experiences are exactly identical.

We focus on the case in which the experiences in the database reflect the objective frequency of events (that of different causes of death in our example). In principle, the database

could be person-specific (e.g., people from New York may hear of fewer experiences of death from tornado than people from Des Moines),[2] and could also be affected by repetition, rehearsal, and prominence of events (e.g., people may hear of more experiences of airplane crashes than of diabetes due to greater news coverage of the former). Furthermore, the database may also include statistical information, as is the case in many experimental settings (Benjamin 2019). The database could also be influenced by selective attention. A past smoker concerned with lung cancer could encode many events of this disease (Schwartzstein 2014). We leave such extensions to future work.

The DM forms beliefs about the relative frequency of two disjoint hypotheses H_1 and H_2, which are subsets of the database E. For instance, the DM may assess the frequency of death by $H_1 =$ "natural disaster" vs. $H_2 =$ "all other causes". These hypotheses partition the subset of causes of death, identified by $f_d = 1$, on the basis of the "natural disaster" feature $f_{d+2} = 1$ vs. $f_{d+2} = 0$. As we describe below, the DM makes his assessment by extracting a sample from his database. Critically, sampling is shaped by similarity and interference, in line with memory research (Kahana 2012; Bordalo et al. 2020). We next present our formalization of similarity.

2.1 Similarity

A symmetric function $S(u, v): E \times E \to [0, \overline{S}]$ measures the similarity between any two experiences u and v in the database. It reaches its maximum $\overline{S}$ at $u = v$. Similarity between two experiences increases in the number of shared features. For instance, a death from a tornado is more similar to that from flooding than either is to death from diabetes, because the former are both caused by a natural disaster rather than an illness. Different features may be differently

[2] Such person-specific databases have many implications that are beyond the scope of this paper. For example, if a person knows many others similar to himself (or simply oversamples them relative to others), he may exaggerate the frequency of traits or opinions that he has. See Mullen et al (1985) for a review of such "false consensus" effects.

weighted, based on their importance or salience. Episodes of a heart attack are similar to each other even if they occur in different contexts. We rely on general intuitions about similarity, not on a particular functional form. A rich literature measures subjective similarity between objects and connects it to observable features (Tversky 1977; Nosofsky 1992; Pantelis et al. 2008).

We define the similarity between two subsets of the database $A \subset E$ and $B \subset E$ to be the average pairwise similarity of their elements,

$$S(A,B) = \sum_{u \in A} \sum_{v \in B} S(u,v) \frac{1}{|A|} \frac{1}{|B|}. \quad (1)$$

$S(A,B)$ is symmetric and increases in feature overlap between members of A and B. The similarity between two disjoint subsets of E is positive if their elements share some features.

We use Equation (1) to define four important objects. The first is the similarity $S(e,H_i)$ between a single experience $e \in E$ and a hypothesis H_i. It increases in the extent to which e shares features with the average member of H_i. Obviously, $e =$ "flood" is similar to $H_1 =$ "natural disaster", while $e =$ "diabetes" is very dissimilar to it. The second object is the self-similarity of hypothesis H_i, $S(H_i,H_i)$. It measures the homogeneity of H_i. Consider $H_1 =$ "natural disaster": A tornado in Tulsa is fairly similar to a tornado in Little Rock, but neither is as similar to an earthquake in California, which reduces the self-similarity of H_1. The third object is "cross-similarity" between hypotheses $S(H_1,H_2)$. In $H_1 =$ "natural disaster", a death from a flood is similar to a death from accidental drowning in H_2, which raises $S(H_1,H_2)$. The fourth and final object is the cross-similarity between H_i and the rest of the database, $\overline{H} = E \backslash H_i \cup H_j$, denoted by $S(H_i,\overline{H})$. When assessing the frequency of different causes of death, $\overline{H}$ is the set of non-lethal events. In $H_1 =$ "natural disaster", a death from flood is similar to the event of surviving a flood

in $\overline{H}$, which raises $S(H_1, \overline{H})$. Throughout, we focus on the case in which a hypothesis is more similar to itself than to other parts of the database, $S(H_i, H_i) \geq \max\{S(H_i, H_j), S(H_i, \overline{H})\}$.[3]

2.2 Memory Sampling

Our formalization of similarity-based sampling and its mapping with beliefs builds on two assumptions. The first formalizes cued recall.

Assumption 1. Cued Recall: *When cued with hypothesis H_i, the probability $r(e, H_i)$ that the DM recalls experience e is proportional to the similarity between e and H_i. That is,*

$$r(e, H_i) = \frac{S(e, H_i)}{\sum_{u \in E} S(u, H_i)}. \quad (2)$$

If $S(u, v)$ is constant, sampling is frequency-based, so $r(e, H_i) = 1/N$. Compared to this benchmark, the numerator of (2) captures the idea that sampling is shaped by similarity to the cue H_i. When thinking about deaths from $H_i =$ "natural disasters", it is relatively easy to recall $e =$ "deaths from floods", due to similarity. The denominator in (2) captures interference: all experiences $u \in E$ compete for retrieval, so they inhibit each other. For instance, when thinking about death by $H_i =$ "natural disasters", the mind may retrieve experiences of different yet frequent lethal events such as $e =$"death from a heart attack".

Interference reflects the fact that we cannot fully control what we recall.[4] It is a well-established regularity in memory research going back to the early 20th century (Jenkins and Dallenbach 1924; McGeoch 1932; Underwood 1957). For example, recall from a target list of

[3] This condition can be violated if H_1 has two opposite clusters and H_2 is in the middle. Consider a database with two generic features, and suppose that the DM assesses hypotheses $H_1 \equiv \{(1,0),(0,1)\}$ and $H_2 \equiv \{(1,1)\}$. Here members of H_1 disagree along all features, while H_2 agrees with one of them, so $S(H_1, H_1) < S(H_1, H_2)$.

[4] In our examples, interference comes from recalling a particular experience, but we do not claim that this process is conscious. Recall failures may manifest as "mental blanks", inability to recall anything when thinking about H_i, or as "intrusions", namely recall of hypothesis-inconsistent experiences $u \notin H_i$.

words suffers intrusions from other lists studied at the same time, particularly for words that are semantically related to the target list, resulting in lower likelihood of retrieval and longer response times (Shiffrin 1970; Lohnas et al. 2015). In the "fan effect", Anderson and Reder (1999) show that concepts associated with more items are more difficult to remember in response to any specific cue.[5] Our application of interference to probability estimates is new. We show that it produces biases linked to the availability and representativeness heuristics.

Our second assumption is that, given the recall probability function $r(e, H_i)$, probability judgments are formed according to the following two-stage sampling process:

Assumption 2. Sampling and Counting

Stage 1: For each hypothesis H_i, the DM samples $T \geq 1$ experiences from E with replacement according to $r(e_k, H_i)$. Denote by R_i the number of successful recalls of experiences in H_i.

Stage 2: The DM estimates the probability of H_i, denoted $\hat{\pi}(H_i)$, as the share of successful recalls of H_i out of all successful recalls of the hypotheses considered:

$$\hat{\pi}(H_i) = \frac{R_i}{R_1 + R_2} \tag{3}$$

Intuitively, the DM draws two random samples, one for each hypothesis.[6] He then counts the number of successes in recalling each H_i, discarding intrusions, and finally estimates the probability of a hypothesis as its relative share of successful recall attempts.

[5] Two approaches have been proposed to account for this evidence: associative models (Anderson and Reder 1999) and inhibition models (Anderson and Spellman 1995). These models do not incorporate interference from irrelevant data. A robust phenomenon related to intrusions is that of "false memories" (Deese 1959; Roediger and McDermott 1995). For example, recall from a list of words that are semantically related suffers intrusions from similar words not on the list, e.g. mis-remembering milk from a list that includes butter, cheese, and white (Brown et al 2000).

[6] Sampling with replacement has two interpretations. The first is that the sample size is small relative to N. The second is that repeated recall of certain events makes them more prominent in mind, affecting beliefs. This is consistent with the fact that unique experienced events such as a stock market crash appear to persistently affect beliefs (Malmendier and Nagel 2011). Sanborn and Chater (2016) allow for more structured Bayesian approaches to frequency-based sampling, such as Markov Chain Monte Carlo. These, however, account neither for the role of similarity nor for systematic violations of consistency such as disjunction and conjunction fallacies.

The assumption that each hypothesis is sampled separately (Stage 1) is especially realistic when the different hypotheses are prominently presented to the DM, which is the case in our experiments. It may be violated if the DM's task is to represent an entire distribution without being cued with specific values, especially if the possible outcomes are many (e.g., the age distribution of deaths), because some outcomes may fail to come to mind, and are not sampled.

The assumption that the DM assesses probabilities by counting "successes" in the drawn samples (Stage 2 above) is realistic in one-shot estimation problems, but may fail in repeated settings, because the DM may learn about the selected nature of the recalled samples. Such learning is unlikely to be perfect, for it itself is subject to memory limitations. Relatedly, the sample size T may be optimized based on the DM's thinking effort. We also assume that when counting "successes", the DM recognizes whether a retrieved memory is consistent with any of the hypotheses. In practice, the DM may have a noisy recollection of a given experience, or may distort recalled experiences self-servingly, as documented in the literature on the hindsight bias (Roese and Vohs 2012). These might be promising extensions of the model.

Finally, in our model the DM forms beliefs by counting the retrieved experiences consistent with each hypothesis and by discarding intrusions. Our model can be extended to account for situations in which the individual assesses a novel scenario, so probability estimates do not involve only counting. In this case, sampling from memory can be accompanied by the simulation of the novel scenario, as documented by a substantial body of work in psychology (Schacter et al. 2007, 2012 and Kahneman Tversky 1981). For example, when assessing the likelihood that a person is a feminist bank-teller, a DM who has never met one may simulate the hypothesis using memories of people who are similar. Bordalo et al (2022) incorporate simulation into a model of memory-based beliefs, and show it helps account for puzzling patterns in beliefs about Covid lethality.

We view our model as the simplest way to introduce similarity into a sampling model. We judge its success by its ability to account for well-known biases, including strong violations of consistency such as partition dependence and the conjunction fallacy, and for recall data.[7]

2.3 Beliefs

To understand the forces shaping belief formation, consider Figure 1.

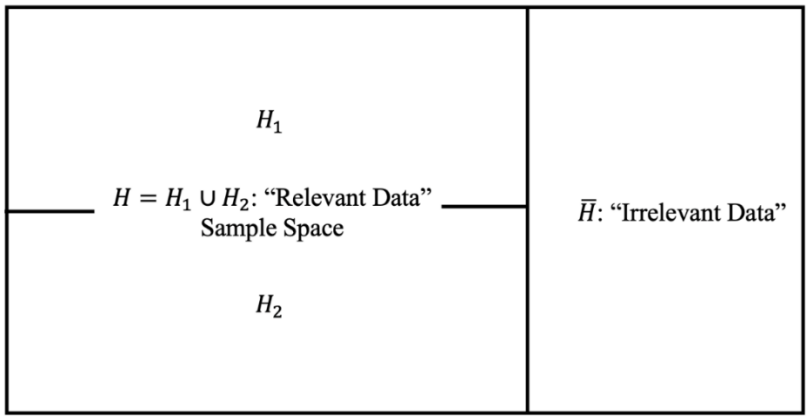

Figure 1: Memory Database and Sample Space

The hypotheses H_1 and H_2 identify three subsets of experiences in E.[8] They also subdivide E into relevant vs. irrelevant experiences. $H = H_1 \cup H_2$ is the set of relevant experiences. In statistics, H is the sample space. The DM forms his subjective beliefs over it. Experiences in $\overline{H} = E \backslash H$, are instead "irrelevant", because they are inconsistent with either hypothesis. When thinking about a hypothesis H_i by sampling in E, similarity causes the DM to focus recall on subset H_i but, by similarity and frequency, sampling may erroneously slip to H_j and $\overline{H}$.

[7] Our model can also be enriched by allowing for: a) sampling to be influenced also by the most recently recalled item, b) the DM to count intrusions from $u \in H_j$, and c) retrieval to be driven by factors other than similarity. For instance, an experience may be more memorable if it is extreme or surprising (Kahneman et al. 1993), or if it is similar to experiences in other contexts, e.g. names of celebrities are more easily remembered (Tversky and Kahneman 1973).
[8] In a slight abuse of notation, we refer to H_i both as a given hypothesis, e.g. "cause of death is flood", and the subset of experiences in E consistent with hypothesis H_i.

Denote by $\pi(H) = |H|/|E|$ the frequency of relevant data in the database and by $\pi(\overline{H}) = |\overline{H}|/|E|$ the frequency of irrelevant data. Denote by $\pi(H_i) = |H_i|/|H|$ the true relative frequency of H_i in the relevant data H, i.e., the correct probability.[9] The total probability that the DM successfully recalls experiences of H_i when thinking about H_i is then given by:

$$r(H_i) = \sum_{e \in H_i} r(e, H_i) = \frac{\sum_{e \in H_i} S(e, H_i)}{\sum_{u \in H_i} S(u, H_i) + \sum_{u \in H_j} S(u, H_i) + \sum_{u \in \overline{H}} S(u, H_i)}$$

$$= \frac{\pi(H_i)\pi(H)}{\pi(H_i)\pi(H) + \frac{S(H_i, H_j)}{S(H_i, H_i)} \cdot \pi(H_j)\pi(H) + \frac{S(H_i, \overline{H})}{S(H_i, H_i)} \cdot \pi(\overline{H})}. \quad (4)$$

In psychology, $r(H_i)$ is known as the retrieval fluency of H_i. It is shaped by two forces: frequency and similarity. It is ceteris paribus easier to recall a more frequent hypothesis, with a higher $\pi(H_i)\pi(H)$. If similarity is constant or if self and cross similarities are equal ($S(H_i, H_i) = S(H_i, H_j) = S(H_i, \overline{H})$), then fluency only depends on frequency, $r(H_i) = \pi(H_i)\pi(H)$.[10]

Similarity shapes sampling in two ways: cueing and interference. First, higher self-similarity $S(H_i, H_i)$ boosts the recall of H_i. This is cueing: thinking about $H_i =$ "flood" cues selective recall of deaths from floods. The retrieval fluency of H_i is then higher than its frequency, especially for unlikely hypotheses. People rarely experience floods and earthquakes compared to heart attacks, so cueing $H_1 =$ "natural disasters" greatly boosts their retrieval.

Second, the DM's ability to successfully sample H_i is hampered by two interference terms in the denominator of (4). The first is "interference from the alternative hypothesis" H_j. When

[9] More precisely, $\pi(H_i)$ is the probability of H_i conditional on the relevant data H. To ease notation, we do not refer to $\pi(H_i)$ as $\pi(H_i|H)$, until we later study conditional beliefs in which the relevant data H is restricted to a subset D.

[10] Similarity does not matter when the DM either samples all data (i.e. $S(H_i, H_j) = S(H_i, H_i) = S(H_i, \overline{H})$), or all relevant data (which occurs when $S(H_i, H_j) = S(H_i, H_i)$, $S(H_i, \overline{H}) = 0$), with equal probability regardless of the cue. In both cases, the expression in Equation (4) becomes proportional to $\pi(H_i)$, so that beliefs are unbiased.

thinking about deaths from H_i = "flood", the mind may retrieve experiences of deaths from causes similar to "flood", such as "accidental drownings" or other natural disasters, that belong to H_j = "other causes of death". In Figure 1, this corresponds to "vertical" intrusions from H_j. Such intrusions are more common when the two hypotheses are more similar, $S(H_i, H_j)$ is higher. The second term is "interference from irrelevant data" $\overline{H}$. When thinking about deaths from H_i = "flood", the mind may retrieve experiences of "surviving floods" that belong to $\overline{H}$ = "non-lethal events". In Figure 1, this corresponds to horizontal intrusions from $\overline{H}$. This effect also hinders sampling of H_i, the more so the higher is cross similarity $S(H_i, \overline{H})$.

We can now describe the probabilistic assessment $\hat{\pi}(H_i)$ in Equation (3). By Assumption 2, the number of successes in recalling each hypothesis H_i follows a binomial distribution: $R_i \sim Bin(T, r(H_i))$. Beliefs $\hat{\pi}(H_i)$ are thus stochastic and characterized as follows.

Proposition 1 *As $T \mapsto \infty$ the distribution of the estimated odds of H_i relative to H_j converges in distribution to a Gaussian with mean and variance:*

$$\mathbb{E}\left[\frac{\hat{\pi}(H_i)}{\hat{\pi}(H_j)}\right] = \frac{r(H_i)}{r(H_j)}. \tag{5}$$

$$\mathbb{V}\left[\frac{\hat{\pi}(H_i)}{\hat{\pi}(H_j)}\right] = \frac{1}{T}\left[\frac{r(H_i)}{r(H_j)}\right]^2\left[\frac{1-r(H_j)}{r(H_j)} + \frac{1-r(H_i)}{r(H_i)}\right]. \tag{6}$$

In (5), the DM attaches a higher probability to hypotheses with relatively high retrieval fluency, as in Tversky and Kahneman's (1973) availability heuristic. If similarity does not drive recall, e.g. $S(u, v)$ is constant, beliefs are frequency based. Thus, average odds in (5) are correct, $r(H_i)/r(H_j) = \pi(H_i)/\pi(H_j)$, but beliefs display noise in (6) due to sampling variance.

When similarity matters, biases arise. We now focus on this case by assuming $S(H_i, H_i) > \max\{S(H_i, H_j), S(H_i, \overline{H})\}$. We study biases in average beliefs, leaving the systematic analysis of memory and noise to future work.[11]

3. Judgment Biases

Section 3.1 shows how similarity affects interference from the alternative hypothesis, yielding biases related to the availability heuristic. Section 3.2 incorporates interference from irrelevant data, and shows that it accounts for the representativeness heuristic. Section 3.3 shows that these two forces can unify over and underreaction of beliefs to data.

3.1 Similarity and Interference from the Alternative Hypothesis

To study interference from the alternative hypothesis, we focus on the case in which the database E coincides with the relevant data for assessing H_1 and H_2 (or equivalently that similarity falls very sharply when moving outside H). In our example, this means that the DM only samples causes of death and there is no intrusion from unrelated events. Furthermore, we assume that T is high enough that average odds are characterized by Equation (5).

Lichtenstein et al (1978) document the overestimation of cued low probability events, such as death from botulism or a flood, and underestimation of cued and likely causes such as heart disease. The average assessed odds in Equation (5) produce this phenomenon.

[11] In general, when two hypotheses are easy to recall—i.e., when both $r(H_1)$ and $r(H_2)$ are high—the variability of beliefs declines, because the DM benefits from a larger sample size. In Appendix B we test this and other predictions about noise in probabilistic assessments.

Proposition 2 *Holding $S(H_i, H_j)$ fixed, the estimate $\hat{\pi}(H_1)$ increases in the objective frequency $\pi(H_1)$. Overestimation, i.e., $\hat{\pi}(H_1) > \pi(H_1)$, occurs if and only if the hypothesis is sufficiently unlikely, $\pi(H_1) < \pi^*$, where threshold π^* is defined by:*

$$\frac{\pi^*}{1-\pi^*} \equiv \frac{1 - \frac{S(H_1, H_2)}{S(H_1, H_1)}}{1 - \frac{S(H_1, H_2)}{S(H_2, H_2)}}. \tag{7}$$

If both hypotheses are equally self-similar, $S(H_1, H_1) = S(H_2, H_2)$, then $\pi^* = 0.5$.

Overestimation of an unlikely hypothesis is due to cued recall of its instances, which occurs because the self-similarity of H_i is higher than its cross similarity with H_j.[12] When thinking about H_1= "floods", the DM selectively retrieves deaths due to floods, oversampling this rare event compared to H_2= "other causes of death". Similarity thus creates insensitivity to frequency, a tendency for beliefs to be smeared toward 50:50.

Kahneman and Tversky's (1979) probability weighting function also features insensitivity to true frequency when weighting objective probabilities.[13] Our model applies to the construction of subjective probabilities, and crucially implies that an unlikely event may be over or underestimated due to similarity-based cueing and interference. In sharp contrast with KT's probability weighting function, in our model unlikely events are prone to be neglected when they are not directly cued. To see this, consider the frequency with which a DM thinking about $H_2 =$ "causes other than flood" samples elements of its subset $H_{21} =$ "tornado" $\subset H_2$ compared to other

[12] The probability of a hypothesis $\pi(H_i)$ can be varied while holding similarities $S(H_i, H_j)$ fixed by keeping the distribution of similarity within hypotheses constant. Formally, this can be obtained by increasing the frequency of each $e_i \in H_i$ proportionally. Equation (1) is in fact homogeneous of degree zero with respect to such change in E.

[13] Recent work founds this function based on the salience of lottery payoff (Bordalo et al. 2012), noisy perception of numerical probabilities (Khaw et al. 2020, Frydman and Jin 2020), and cognitive uncertainty (Enke and Graeber 2019). In the appendix, we show that recall is strongly correlated with a measure of subjective uncertainty.

elements in H_2. Such relative frequency, given by $r(H_{21}, H_2)/r(H_2)$, is the belief the agent implicitly puts on tornadoes compared to H_2.

Corollary 1 *H_{21} is undersampled compared to its true frequency in H_2 if and only if $S(H_{21}, H_2) < S(H_2, H_2)$. Holding fixed $S(H_{21}, H_2)$ and $S(H_2, H_2)$, there is a threshold π^{**} such that H_{21} is underestimated compared to H_2 if and only if it is unlikely enough, $\pi(H_{21}) < \pi^{**}$.*

Roughly speaking, a non-cued event H_{21} is neglected if it is less similar to the cued hypothesis H_2 to which it belongs than to the average member of H_2, $S(H_{21}, H_2) < S(H_2, H_2)$. This depends, among other things, on the event's frequency: the rarer is H_{21}, the more atypical it is of the cued H_2 and hence the more dissimilar it is to the latter. When thinking about $H_2 =$ "causes other than flood", we may recall the likely "heart attack", not the unlikely "tornado".

Proposition 1 and Corollary 1 also explain why, when thinking about $H =$ "murders in Michigan", people tend to neglect the non-cued subset $H' =$ "murders in Detroit", because Detroit is dissimilar to the rest of Michigan. In contrast, people tend to overestimate $H' =$ "murders in Detroit" when explicitly cued (Kahneman and Frederick 2002).

A second implication of similarity is that an event can be overestimated if homogeneous but underestimated if heterogeneous, regardless of its likelihood.

Corollary 2. *Holding fixed $\pi(H_1)$, as the events in H_1 become more homogeneous, i.e. $S(H_1, H_1)$ increases, the probability assessment $\hat{\pi}(H_1)$ increases. If $S(H_1, H_1) > S(H_2, H_2)$, the threshold of Proposition 2 satisfies $\pi^* > 0.5$, and H can be overestimated even if it is likely.*

When H_1 becomes more self-similar, it is easier to recall. As a result, it is less likely that, when thinking about it, the mind slips to its alternative hypothesis H_2. According to Equation (5), this increases the estimation of H_1, even if its objective probability stays constant.

This result indicates that cued unlikely events are prone to overestimation because they are often more self-similar than their alternative. When cued by H_1 = "flood", it is easy to imagine instances of this disaster, because they are similar to each other. By contrast, the alternative $H_2 =$ "causes other than flood" is very heterogeneous, and hence hard to imagine. This creates strong interference for H_2, hindering its assessment.

Tversky and Kahneman (1983) asked one group of subjects to assess the share of $H_1 =$ "words ending with _n_" in a certain text. Another group of subjects was asked to assess the probability of $H_{11} =$ "words ending with _ing". Remarkably, subjects attached a lower probability to H_1 than to H_{11}, despite the fact that H_{11} is a subset of H_1. Similarity accounts for this phenomenon: instances of $H_{11} =$ "words ending with _ing" share many features, such as being gerunds, denoting similar activities, etc, which brings many examples to mind. In contrast, $H_1 =$ "words ending with _n_" includes many words which do not share these features (and which often do not share many features with each other). This reduction in self-similarity makes it harder to recall words in H_1, causing its underestimation compared to its subset H_{11}.[14]

A third implication, following from Corollaries 1 and 2, is partition dependence. The total likelihood of death is estimated to be lower for "natural causes" than for "cancer, heart attack or other natural causes" (Tversky and Koehler 1994). Many famous studies document this phenomenon (Benjamin 2019).[15] In our model, it arises because partitioning a hypothesis into more specific sub-events increases its overall self-similarity, reducing interference. To see this,

[14] To check this intuition, we ran a simple online survey. Respondents indeed rate randomly generated groups of "_ing" words as being more similar to each other than groups of _n_ words. Results are available on request.

[15] For example, Fischoff et al. (1978) famously show that when assessing the cause of a car's failure to start, mechanics judge "ignition" more likely when alternative causes were partitioned into "ignition", "fuel", "other". Sloman et al. (2004) show, in contrast, that death by "pneumonia, diabetes, cirrhosis or any other disease" is estimated to be less likely than death by "any disease". This is consistent with an extension of our model in which atypical cues such as "cirrhosis" focus attention on a narrow subset, interfering with the retrieval of more common diseases. A similar pattern occurs in free recall tasks (Slamecka 1968; Sanborn and Chater 2016).

suppose that the alternative hypothesis H_2 is explicitly partitioned into H_{21} and H_{22}. The subsets are equally: i) likely, $\pi(H_{21}) = \pi(H_{22})$, ii) self-similar, $S(H_{21}, H_{21}) = S(H_{22}, H_{22})$, and iii) cross-similar, $S(H_{21}, H_1) = S(H_{22}, H_1)$.[16] We then obtain:

Proposition 3 *Partitioning the alternative hypothesis H_2 into H_{21} and H_{22} is equivalent to increasing the self-similarity of H_2 if and only if:*

$$S(H_{21}, H_{21}) > S(H_{21}, H_{22}). \tag{8}$$

In this case, partitioning H_2 reduces $\hat{\pi}(H_1)$, and the more so the higher is $\frac{S(H_{21},H_{21})}{S(H_{21},H_{22})}$.

The assessment of a given hypothesis H_1 = "flood" is reduced when its alternative is specified as H_{21} = "natural causes" and H_{22} = "non-natural causes other than flood", compared to when it is specified as H_2 = "causes other than flood". Cueing H_{21} and H_{22} fosters retrieval of alternatives to flood, which reduces the assessment of H_1 = "flood". Tversky and Koehler's (1994) "Support Theory" offers an explanation based on the idea that people evaluate events using a sub-additive "support function". In our model, partition dependence comes from similarity in recall.

In sum, with similarity-based sampling, the DM evaluates a hypothesis by retrieving instances of it, but in doing so finds it hard not to think about the alternative hypothesis. Such interference reconciles well-known biases such as smearing of probability judgments toward 50:50, underestimation of rare events that are not cued, and availability effects in which the similarity structure of hypotheses and their description affect beliefs. Appendix A0 summarizes the main biases explained by our model and the required conditions on the similarity function.

3.2 Biases due to Interference from Irrelevant Experiences

[16] These conditions nest three hypotheses (H_1, H_{21}, H_{22}) in the binary hypotheses case, connecting to Proposition 2.

Until now, we ruled out interference from irrelevant data by assuming that the database E coincides with the relevant data $H = H_1 \cup H_2$. Suppose, however, that the DM must condition H_1 and H_2 on data D, which identifies a subset $D \subset H$. For concreteness, the DM assesses deaths by H_1 = "accident" vs. H_2 = "sickness" in the specific group of D = "young". The DM samples the events $H_1 \cap D$ = "accidents among the young" and $H_2 \cap D$ = "sickness among the young" using the retrieval fluencies $r(H_1 \cap D)$ and $r(H_2 \cap D)$.[17] Critically, now irrelevant experiences from $\overline{D} = E \backslash D$ can interfere, in our example those of $\overline{D}$ = "older" people. We show that this kind of interference produces effects typically explained using the representativeness heuristic.

To visualize interference from $\overline{D}$, Figure 2 depicts the database E, where the size of each region roughly corresponds to true frequencies.

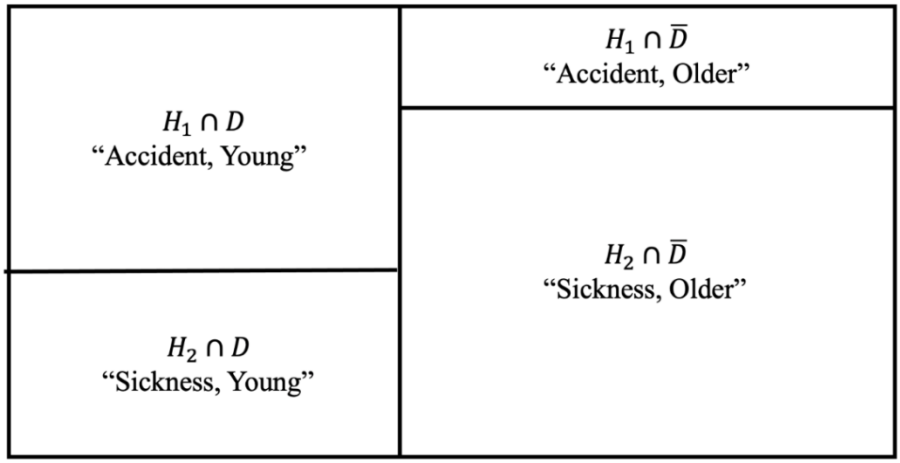

Figure 2: Visualizing conditional assessments

When thinking about $H_1 \cap D =$ "accident among the young", two kinds of interference are at work. First, as in our prior analysis, there is vertical intrusion of memories of young people dying from sickness (i.e. from $H_2 \cap D$), due to similarity among young people. But second, there

[17] These retrieval fluencies are still defined by Equation (4) with the change in notation $H = D$ and $\overline{H} = \overline{D}$.

is horizontal intrusion of irrelevant experiences of older people dying from accidents (i.e., from $H_1 \cap \overline{D}$), due to the similarity along the $H_1 =$ "accident" dimension. Similarly, when thinking about $H_2 \cap D$ = "sickness among the young", the DM faces vertical intrusion from "accidents among the young" and horizontal intrusions from the irrelevant "sickness among the older".

Interference from $\overline{D}$ may affect one hypothesis more than the other. In our example, the deaths of older people interfere more with thinking about "sickness" because the bulk of the elderly die from sickness, not from accidents. Thus, when thinking about young people dying from sickness many old people dying from sickness intrude, while intrusions are few when thinking about accidents. This effect can cause overestimation of $H_1 \cap D =$ "accident among the young".

Formally, suppose that there are only two features (in our case the cause of death, accident vs. sickness, and age, young vs older). The DM assesses the distribution of the first feature (cause of death) conditional on a value of the other (young). Suppose furthermore that similarity takes the functional form: $S(e, e') = \delta^{\sum_i |f_i - f_i'|}$, so it decreases by a factor of δ for each differing feature. We denote the *conditional* probability estimate obtained using Equation (4) by $\hat{\pi}(H_i|D)$ and we compare it to the true conditional probability $\pi(H_i|D)$.

Proposition 4. *For $\delta < 1$, the DM overestimates the probability of H_1 conditional on D, $\hat{\pi}(H_1|D) > \pi(H_1|D)$, if and only if:*

$$\pi(H_1|D)\pi(D) + \delta\pi\left(H_1\middle|\overline{D}\right)\pi\left(\overline{D}\right) < \frac{\pi(D) + \delta\pi\left(\overline{D}\right)}{2}. \quad (9)$$

The first term on the left-hand side is standard: overestimation is more likely when the true conditional probability $\pi(H_1|D)$ is low, in line with Section 3.1. The second term is new: the conditional hypothesis is overestimated also if its frequency in the irrelevant data, $\pi\left(H_1\middle|\overline{D}\right)$, is

low. In this case, H_1 is less similar to the irrelevant data $\overline{D}$ than H_2. Thus, H_1 faces less interference than H_2 from irrelevant data, which promotes overestimation of the former.

Consider this effect in Figure 2. H_1= "accident" is a common cause of death for the young ($\pi(H_1|D)$ is high), so interference from the alternative hypothesis promotes its underestimation. At the same time, when considering young people dying from sickness, many instances of the old dying from sickness intrude ($\pi(H_1|\overline{D})$ is low). This can cause overestimation of H_1 = "accident" for the young, even if for them it is the more likely cause of death.

Intrusion of irrelevant data sheds light on Kahneman and Tversky's (1973) representativeness heuristic, including the so-called conjunction fallacy in the Linda problem. Subjects are told that Linda was an activist in college, so D = "activist". Some are then asked the probability that she is currently a H_1 = "bank teller", others that she is a H_{11} = "feminist bank teller". Strikingly, feminist bank teller is rated likelier than bank teller even though $H_{11} \subset H_1$. According to Proposition 4, this occurs for two reasons. First and foremost, H_{11} = "feminist bank teller" is much less similar to the group of $\overline{D}$ = "non-activists" than H_1 = "bank teller". Intuitively, among "non-activists" there are many fewer feminist bank tellers than bank tellers, $\pi\left(H_1|\overline{D}\right) > \pi\left(H_{11}|\overline{D}\right)$. Thus, H_{11} = "feminist bank teller" faces less interference from irrelevant data than H_1 = "bank teller", which promotes overestimation of H_{11}. Second, H_1 = "bank teller" is likelier than H_{11} = "feminist bank teller", $\pi(H_1|D) > \pi(H_{11}|D)$, which also promotes underestimation of H_1 relative to H_{11}. Both effects create this conjunction fallacy.

Tversky and Kahneman (1983) define representativeness as follows: "an attribute is representative of a class if it is very diagnostic; that is, the relative frequency of this attribute is much higher in that class than in a relevant reference class." Gennaioli and Shleifer (2010) formalize this idea by assuming that the conditional probability $\pi(H|D)$ is overestimated if the

likelihood ratio $\frac{\pi(H|D)}{\pi(H|\overline{D})}$ is high. The irrelevant data $\overline{D}$ captures the "reference class" in Tversky and Kahneman's definition. Bordalo et al. (2016) use this idea to model social stereotypes, while Bordalo et al. (2018) use it to model diagnostic expectations.

Intrusion from irrelevant data provides a foundation for representativeness and the reference class $\overline{D}$ in a way that squares with Kahneman and Tversky's (1973) broad intuition that similarity judgments affect beliefs. When $\pi(H_1|\overline{D})$ is low, H_1 is dissimilar from the irrelevant data $\overline{D}$. As a result, H_1 suffers less interference from $\overline{D}$ than does H_2, so experiences in $H_1 \cap D$ are easier to retrieve causing overestimation of $\pi(H_1|D)$. In Section 5.3 we show that this mechanism explains "kernel of truth" stereotypes (BCGS 2016) and yields new predictions.

One advantage of our approach is to identify limits to representativeness, which are due to strong intrusion from the alternative hypothesis. We now show how the interaction between these forces throws new light on the conflicting evidence of over and underreaction.

3.3 Underreaction and overreaction to data

Work from the lab and the field documents conflicting distortions in belief updating. There is evidence that people overestimate the probability of events in light of data informative about them, a finding often explained by the representativeness heuristic (Kahneman and Tversky 1973). There is also evidence of underestimation in similar situations, often explained with inattention (Sims 2003; Coibion and Gorodnichenko 2012; Gabaix 2019). Memory helps unify this evidence, yielding conditions as to when either phenomenon should occur.

To connect to this debate, we define over and underreaction. We say that the DM overreacts to data D if: 1) D is objectively informative about a hypothesis H_i, i.e. $\pi(H_i|D) > \pi(H_i)$, and 2) the DM overestimates that hypothesis, i.e. $\hat{\pi}(H_i|D) > \pi(H_i|D)$. The DM underreacts otherwise.

This definition captures the intuition of overreaction in many real-world settings in which the DM's prior belief and the likelihood function are unavailable, as with stereotypes (red haired Irish) or the Linda problem.[18] Proposition 4 implies the following result.

Corollary 3. *Suppose that D is informative about* H_1*. If the true probability* $\pi(H_1|D)$ *is higher than a threshold* $\bar{\pi} > 0.5$*, the DM underreacts to D. If* $\pi(H_1|D) < \bar{\pi}$*, the DM overreacts to D if* $\pi(H_1|D)$ *or* $\pi(H_1|\overline{D})$ *are sufficiently low, and underreacts to D otherwise.*

In Figure 3, the data is informative about H_1 in the region above the 45° line.[19] In region A, H_1 is overestimated, while in region B it is underestimated, as per Equation (9).

Consider different cases, starting from the two most extreme ones. In the lower left corner of region A, overreaction is strong. Here interference is low both from the irrelevant data ($\pi(H_1|\overline{D})$ is low) and from the alternative hypothesis ($\pi(H_1|D)$ is low). The resulting overreaction takes the form of base rate neglect. In Tversky and Kahneman (1974), people overestimate the chances that Steve, a "shy and withdrawn person with a passion for detail," is a librarian rather than a farmer, even though farming is a much more common occupation, especially among men. Overreaction occurs because librarians are relatively rare ($\pi(H_1|D)$ is low) and because many farmers are neither shy nor have a passion for detail, so farmers are more similar to irrelevant data than librarians ($\pi(H_1|\overline{D})$ is low, which implies $\pi(H_2|\overline{D})$ is high).

At the other extreme, in the upper part of Figure 3, when $\pi(H_1|D) > \bar{\pi}$, interference from the alternative hypothesis is very strong. Here underreaction occurs, and takes the form of general

[18] In Appendix A4, we show that our definition is equivalent to saying that the DM overreacts if and only if an upward revision of his belief in response to the data ($\hat{\pi}(H_i|D) > \hat{\pi}(H_i)$) is associated with an overestimation, or negative prediction error ($\hat{\pi}(H_i|D) > \pi(H_i|D)$). This criterion is often used to detect over and underreaction in the field, using data on revisions of expectations (Coibion and Gordnichenko 2012, and Bordalo et al. 2019). When priors and likelihoods are available, under and overreaction is often defined in terms of *sensitivities* rather than *levels*, as is done in Grether (1980), i.e. in terms of the difference between the elicited prior and posterior beliefs.

[19] This is equivalent to $\pi(H_1|D) > \pi(H_1)$. The characterization is identical for H_2 when $\pi(H_1|D) < \pi(H_1)$. In fact, overreaction is given by either $\hat{\pi}(H_1|D) < \pi(H_1|D) < \pi(H_1)$ or $\hat{\pi}(H_1|D) > \pi(H_1|D) > \pi(H_1)$.

conservatism and aversion to extreme beliefs (Griffin and Tversky 1992; Benjamin 2019). Even though the data point to H_1, cueing the unlikely alternative H_2 causes its overestimation.

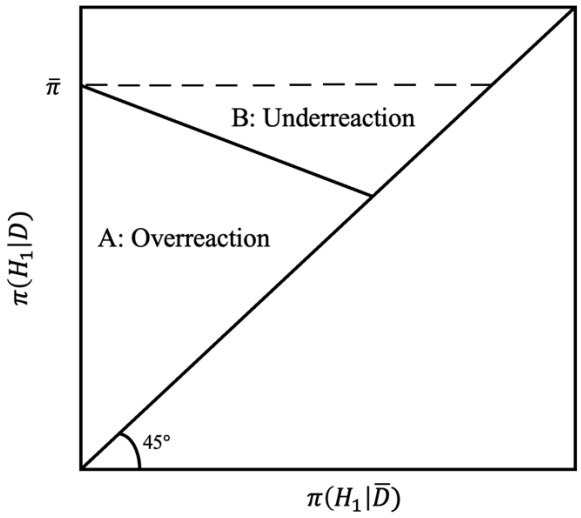

Figure 3. Condition for underreaction and overreaction to data

Notes: This figure depicts the region of $\left(\pi(H_1|D),\ \pi(H_2|\ \bar{D})\right)$ where the agent overreacts or underreacts to data D, where D is diagnostic of H ($\pi(H_1|D) > \pi(H_1|\bar{D})$). Region A corresponds to overreaction, region B to underreaction.

In the intermediate region of Figure 3, H_1 is moderately likely. Whether over or under reaction prevails depends on the signal's strength. Overreaction occurs when the signal is strong, in the upper part of region A. An investor may overestimate the probability that a firm has H_1 = "strong fundamentals" if it has experienced D = "rapid earnings growth." This occurs provided firms with strong fundamentals rarely exhibit lackluster growth, i.e., $\pi(H_1|\bar{D})$ is low. In this case, rapid earnings growth makes it easy to think about strong fundamentals. It is instead harder to think about H_2 = "weak fundamentals," because these firms often produce $\bar{D}$ = "not rapid earnings growth". The DM overreacts because H_1, although likely, faces much less interference from irrelevant data than H_2. This logic offers a micro-foundation for diagnostic expectations.

On the other hand, if the hypothesis is moderately likely but the signal is weak, there is underreaction, due to high interference from irrelevant data $\pi(H_1|\overline{D})$. Suppose that the DM evaluates a firm and the data is not "rapid growth" but rather $D =$ "positive earnings surprise". Even firms with strong fundamentals may well have negative earnings surprises, so $\pi(H_1|\overline{D})$ is higher than in the previous example. This creates interference for H_1, potentially causing its underestimation and underreaction. If D points to a fairly likely hypothesis, beliefs underreact to weakly diagnostic data and overreact to strongly diagnostic data.

To summarize, similarity and interference in human recall naturally account for a range of well-documented biases due to Kahneman and Tversky's availability and representativeness heuristics, and shed light on conflicting evidence on under and overreaction.

4. Experiments

We assess our key predictions in two "pure recall" experiments in which we modulate similarity and interference by exogenously varying subjects' databases and cues. Experiment 1 studies the role of interference from the alternative hypothesis. Experiment 2 additionally studies interference from irrelevant data. In both experiments subjects first go through a controlled set of experiences in which they see a series of images, and then make a probabilistic assessment about them. To do so, they only need to recall the images they saw earlier. Relative to conventional designs, which provide subjects with statistical information (e.g., Edwards 1968; Enke and Graeber 2019) or ask hypothetical questions about naturalistic situations (Kahneman and Tversky 1973), our approach i) allows us to control the memory database, ii) avoids anchoring to given numerical probabilities, and iii) enables us to measure recall of specific experiences, and thus to assess whether recall and probability estimates go hand in hand.

Subjects were recruited from Bocconi University undergraduates on the experimental economics email list. They could participate in both experiments, which occurred four months apart, and completed the experiments remotely due to Covid restrictions. They earned a 4 euro Amazon gift card, plus a bonus if their answer to one randomly chosen question was accurate.[20] Experiments were pre-registered, including hypotheses and sample sizes, on the AEA RCT Registry, with ID AEARCTR-0006676. Appendix B provides more details about both surveys.

4.1 Experiment 1: Testing Interference from the Alternative Hypothesis

This experiment tests three key implication of interference from the alternative hypothesis.

Prediction 1: Memory creates a tendency to overestimate cued unlikely hypotheses, and overestimation is stronger for rarer hypotheses. (Proposition 2)

Prediction 2: Holding objective frequencies constant, the assessed probability of a hypothesis increases when its alternative is more heterogeneous/less self-similar. (Corollary 2)

Prediction 3: Holding objective frequencies constant, the assessed probability of a hypothesis decreases if its alternative is partitioned into two more self-similar subsets. (Proposition 3)

Participants are told that they will see 40 words, one by one in a random order. They are told that they will then be asked questions about the words and that answering them correctly will increase their chances of winning a bonus payment.[21] They are not told what the questions will be. In all treatments, some of the words are animals and some are not, though participants are not informed of this ahead of time. In three treatments they are then asked the following question:

[20] If the chosen question was a probability estimate, they earned 2 euros if their answer was within 5 percentage points of the truth. If it was a free recall task, each correctly/incorrectly recalled word increased/decreased subjects' chance of winning the 2 euro bonus by 10 percentage point (bounded by 0 and 1). The bonus provides easy to understand incentives compared to other schemes such as binarized scoring rules, which distort truth telling (Danz et al., 2020).

[21] Participants answer three comprehension questions that ask them to re-describe each piece of the instructions. Eighty-nine percent of respondents answer all three questions correctly. The results that we present are unchanged if we exclude the 11% who answered at least one question incorrectly.

"Suppose the computer randomly chose a word from the words you just saw. What is the percent chance that it is....

an animal? ____%

anything else? ____%"

The two probabilities must add up to 100%. Afterward, participants are asked to list up to 15 animals and then up to 15 other words that they remember seeing. In all treatments, all exhibited words are relevant to answering the question. Thus, there is no interference from irrelevant data.

The four treatments to test Predictions 1-3 are:

T1: 20% of the words are animals. 80% of the words are names (half male and half female).

T2: 40% of the words are animals. 60% of the words are names (half male and half female).

T3: 40% of the words are animals. The remaining words do not belong to any common category, and hence are very dissimilar to one another.

T4: The distribution of words is as in T2, but subjects are asked about the probability of animals, men's names, and women's names. Assessments must add up to 100%.

These experimental treatments are summarized in Table 1.[22]

Comparing T1 and T2 offers a test for Prediction 1: we expect overestimation of $\hat{\pi}(animal)$ especially in T1, when animals are objectively rarer. By comparing T2 and T3 we can

[22] Three remarks on our experiments. First, Heterogeneous words in T3 were chosen using a random word generator, eliminating words that we deemed too similar to each other (e.g., Mayor, Elected, Town). Second, in the recall task in T4, participants are asked to list up to 15 examples each of animals, men's names, and women's names. Third, In addition to treatments T1-T4, we ran a treatment T5, where we replaced women's names in T1 with ocean animals (e.g., "Shark", "Starfish", "Dolphin", etc.). Participants are then asked the probability of "Land Animals" (in T1, all animals are land animals) and "Anything else." In the recall task, participants are asked to list examples of "land animals" and "other words" that they recall seeing. By increasing cross similarity $S(H_1, H_2)$, this treatment should exert an ambiguous effect on assessments, but it should reduce the ability to recall examples of H_1= "Land Animals". Though the recall data appear consistent with this hypothesis, there was an unexpected confusion about what counted as a land animal: over a quarter of respondents list at least one ocean animal in the free recall task when asked to list land animals. For comparison, no respondents list names when prompted to recall animals in T1. We are therefore less confident in the data from T5 and exclude it from the main analysis. The Appendix describes this issue and the results from this treatment in greater detail. Our main results in Section 4.5 and 5 hold qualitatively if we include T5.

test Prediction 2: compared to T2, $\hat{\pi}(animal)$ should be higher in T3, because the alternative hypothesis (non-animals) is very heterogeneous. By comparing T4 and T2 we can test Prediction 3: in T4 the alternative hypothesis is split into two more self-similar sub-hypotheses (men's names and women's names), so $\hat{\pi}(animal)$ should be lower than in T2. Lastly, the treatment effects on the recall task should mirror those on $\hat{\pi}(animal)$. This is not necessarily due to a causal effect of recalled examples on probability estimates, as both outcomes may be products of retrieval fluency.

Table 1: Treatments in Experiment 1

Treatment	Sample Size	Distribution of Images	Examples	Elicited Belief
T1	$N = 244$	20% Animals, 80% Names	Lion, John, Moose, Rat, Margaret, Deer, Edward, Nancy, Wolf…	P(Animal) vs P(Other)
T2	$N = 244$	40% Animals, 60% Names	Paul, John, Moose, Rat, Margaret, Deer, Laura, Nancy, Edward…	P(Animal) vs P(Other)
T3	$N = 241$	40% Animals, 60% Heterogeneous	Lion, Sled, Moose, Rat, Pure, Deer, Half, Good, Wolf…	P(Animal) vs P(Other)
T4	$N = 234$	40% Animals, 60% Names	Lion, John, Moose, Rat, Margaret, Deer, Edward, Nancy, Wolf…	P(Animal) vs P(Men) vs P(Women)

4.2 Experiment 1 Results

Figure 4 shows the treatment effects. Panel A reports the under or overestimation of $\hat{\pi}(animal)$ compared to the truth. Panel B reports the share of animals among all recalled examples.[23]

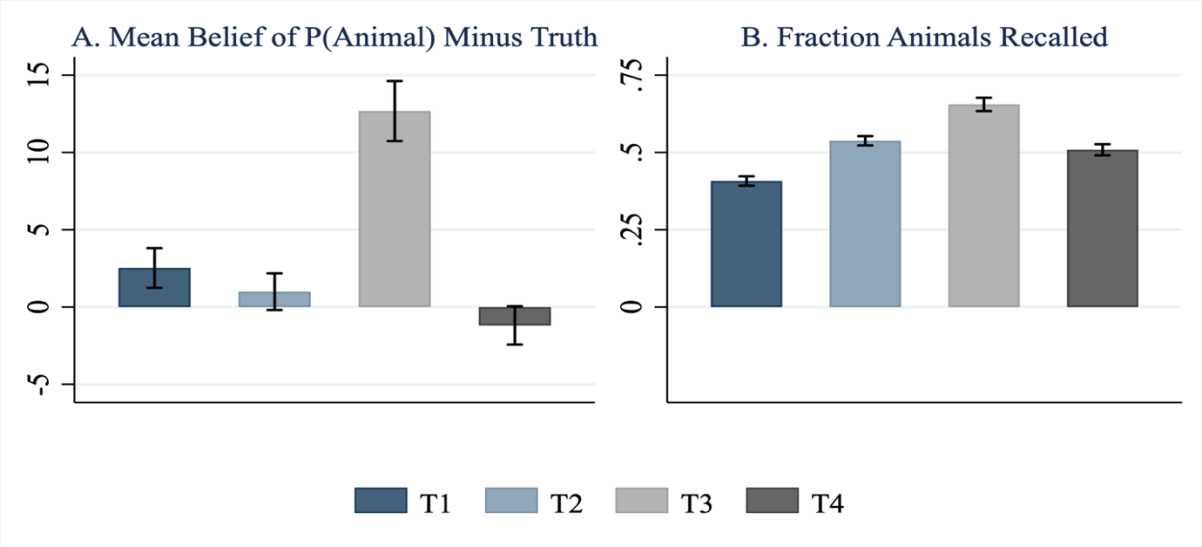

Figure 4: Results from Experiment 1

Notes: This figure shows mean belief of the probability of animals (Panel A) and the mean fraction of recalled words that were animals (Panel B) in Experiment 1. Bands show 95% confidence intervals. The distribution of words for each treatment are: *T1*: 20% Animals, 40% Men's Names, 40% Women's Names, *T2*: 40% Animals, 30% Men's Names, 30% Women's Names, *T3*: 40% Animals, 60% Heterogeneous words, *T4*: 40% Animals, 30% Men's Names, 30% Women's Names.

Consistent with Prediction 1, there is a tendency to overestimate $\hat{\pi}(animal)$, especially in T1 where animals are only 20% of words: overestimation of animals (that is, mean belief minus truth) is 2.5 percentage points (pp) in T1 and 1 pp in T2, and only the former is significantly different from zero at conventional levels ($p < 0.01$ and $p = 0.10$, respectively). Also, the overestimation in T1 is marginally statistically different from that in T2 ($p = 0.09$).

The result of T3 is striking: consistent with Prediction 2, when we replace people's names with heterogeneous words while keeping the true frequency of "animal" constant at 40%, the

[23]Throughout the analysis to follow, we look at treatment effects on the number of *correctly* recalled words. About 18% of the answers to recall questions (which were free text entry) are not in fact words that were shown to participants, or are words corresponding to other hypotheses. Unless otherwise noted, results look very similar if we instead use the number of recall *entries* (regardless of whether they were correct or incorrect) for a category.

overestimation of $\hat{\pi}(animal)$ increases from 1 pp in T2 to 12.7 pp in T3 ($p < 0.01$). Thus, overestimation depends not only on actual frequency but also on how self-similar the alternative hypothesis is. This effect can dominate attenuation to 50:50: in T3, $\hat{\pi}(animal)$ overshoots 50% ($p = 0.01$). The role of similarity in recall emerges as a powerful force in probabilistic assessments.

Finally, when partitioning "non-animals" into the finer sub-hypotheses "men's names" and "women's names" in T4, the assessment $\hat{\pi}(animal)$ falls by 2.1 percentage points compared to T2 ($p = 0.013$), and is even underestimated relative to the truth ($p = 0.060$). Similarity-based recall implies that the more specific cues in the partition of H_2 can turn overestimation of an unlikely hypothesis H_1 (as in T2) into its underestimation (in T4).

The treatment effects on recall, shown in Panel B of Figure 4, mirror those on beliefs. Significantly fewer (40% vs 54%) of recalled words are animals in T1 compared to T2 ($p < 0.01$), because there are objectively fewer animal words in the former. In T3, where non-animals are heterogeneous words, recall of animals jumps to 66%, significantly higher than in T2 ($p < 0.01$). Finally, in T4, where men's and women's names are separated out, significantly fewer recalled words are animals (50%) than in T2 ($p = 0.02$).[24]

One might worry that our results are driven by differences in attention or encoding of words across treatments, rather than by the intended retrieval mechanism. However, until the words are presented, all treatments are identical to participants, so our treatment effects cannot be due to differences in what participants expected to see or be asked. In addition, in treatments T1, T2 and

[24]While the treatment effects on recall and probability are aligned qualitatively, the exact magnitudes need not align. Indeed, the magnitude of the effect on recall seems to be greater than the effect on probability estimation. In general, the explicitly recalled samples and the internal recall fluency used in probability judgments may not be the same.

T4 we use the same words, and in T4 the distribution of words is identical to T2, only the question-cue is changed. Differences in cueing/retrieval are thus likely driving our results.[25]

We conclude the analysis of Experiment 1 by looking at the link between beliefs and recall at the individual level.

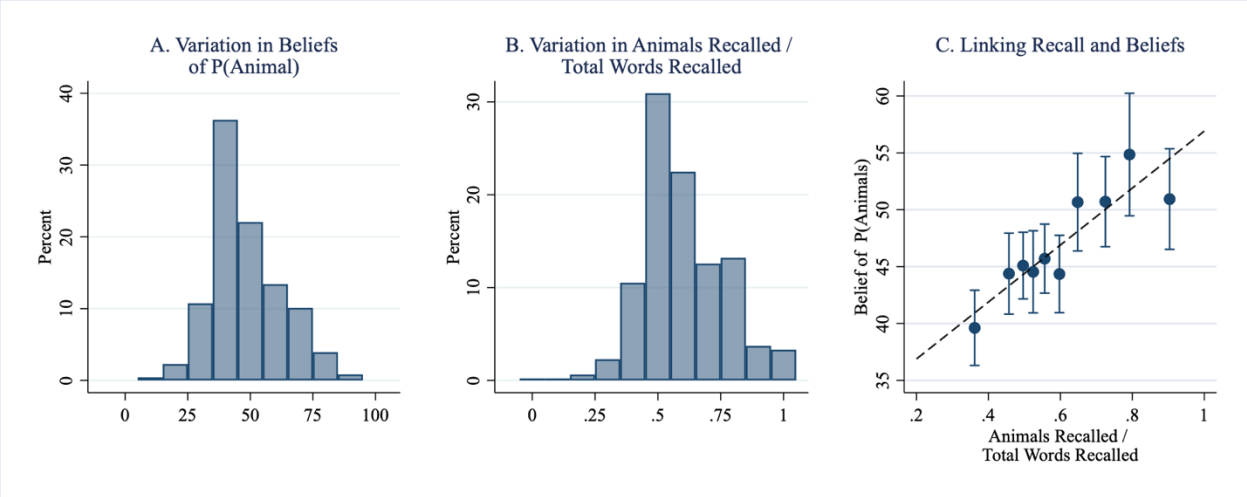

Figure 5: The relationship between recall of examples and beliefs

Notes: Panel A shows the distribution of beliefs about the probability of animals in Experiment 1. Panel B shows the distribution of the number of animals recalled divided by the total number of words recalled. Panel C bins the data by deciles of animals recalled divided by total number of recalled words (x-axis) and shows mean beliefs of the probability of animals (y-axis) The dashed line shows the OLS line of best fit. Bands show 95% confidence intervals. All panels restrict the data to T2 and T3.

Figure 5 pools the T2 and T3 treatments, where the true distribution of words includes 40% animals and 60% non-animals (results look similar if we include the other treatments). There is substantial heterogeneity in beliefs (Panel A) and in the fraction of recalled words that are animals (Panel B). Crucially, beliefs and recall are highly correlated: respondents who recall relatively more animals estimate the probability of drawing an animal to be higher, adding credence to our

[25] Consistent with respondents' answers, we do find evidence in the recall data for primacy effects whereby words that were (randomly) presented earlier in the sequence are more likely to be recalled. See Appendix C for more details, including robustness to controlling for such effects.

interpretation that beliefs and free recall are both dependent on retrieval fluency (Panel C).[26,27] In the appendix, we also show that, in line with Equation (6), participants who recall more words (a proxy for T) also have less variable beliefs (Kahneman et al. 2021).

4.3 Experiment 2: Interference from Irrelevant Data

We designed Experiment 2 to test the implications of interference from irrelevant data. Participants are told that they will be shown 40 images, each of which is either a word or a number, and either orange or blue. Participants were not told how many images would be of each color, but in all treatments 20 images are orange, and 20 are blue.[28] Participants are told that their bonus will depend on their answer to questions about the images, but are not told what the question will be. After seeing the images, one-by-one in a random order, participants in all treatments are asked: "Suppose the computer randomly chose an image from the images you just saw. It is *orange*. What is the percent chance that it is a word?"

Participants must thus assess the probability $\hat{\pi}(w|o)$ that an image is a word conditional on the data that it is orange. Participants answer by clicking on a slider that ranges from 0% to 100%.[29] They are then asked to list up to 10 orange words that they recall seeing.

In this experiment, a subset of experiences – blue words and blue numbers – are irrelevant for assessing the distribution of orange images. Crucially, as subjects try to recall orange words

[26] The simplest interpretation of Figure 8 is that the experiences randomly recalled when making probability judgments later cue recall of the same experiences when asked to list exemplars. If so, probability assessments and free recall are correlated because the distribution of recalled experiences are similar in the two cases. Alternatively, there may be individual-level differences in the subjective similarity of objects to a given cue, making it easier for some subjects to recall certain exemplars for a given category than for others, which will also be reflected in probability assessments.
[27] The correlation is not causal, but it is also not mechanical: subjects are separately asked the percent chance that a randomly chosen word is an animal and then to recall up to 15 examples of each hypothesis.
[28] All words in this experiment are related to time (e.g., "Second", "Week", "Duration", etc.), although we do not ask about word categories so this does not matter for the analysis.
[29] The slider begins with no default, so that participants have to click somewhere on the slider and then move the drag-able icon (that then appears where they first click) to indicate their answer.

(numbers), the irrelevant blue words (numbers) may come to mind and interfere. Of course, interference from the alternative hypothesis is also at play: when thinking about orange words, orange numbers may also come to mind, causing smearing toward 50:50.

To identify interference from irrelevant data, we fix the share of orange images that are words, $\pi(w|o)$, and vary the share of blue images that are words, $\pi(w|b)$. In *Low* treatments no blue image is a word, $\pi(w|b) = 0$, in *High* treatments all blue images are words, $\pi(w|b) = 1$, and in *Middle* treatments either half or 30% of blue images are words, $\pi(w|b) = 0.5, 0.3$. Our model predicts that higher $\pi(w|b)$ should reduce both the estimated $\pi(w|o)$ and recall of orange words.

We study how interference from irrelevant data interacts with that from the alternative hypotheses. To do so, we vary the share of orange images that are words, $\pi(w|o)$, i.e. the correct answer. In *Neutral* treatments the true answer is 50%, $\pi(w|o) = 0.5$. In *Intermediate* treatments the true answer is 55%, $\pi(w|o) = 0.55$. In *Common* treatments the true answer is 70%, $\pi(w|o) = 0.7$. Due to growing interference from the alternative hypothesis, treatments with higher $\pi(w|o)$ should see a stronger tendency toward underestimating orange words, potentially even if there is very little interference from irrelevant data, namely even if $\pi(w|b)$ is very low.

In the same experimental setting, BCGSS (2021) show that increasing the association of irrelevant data with a category causes a lower probability estimate for that category, in line with our treatments varying $\pi(w|b)$ here. The novelty of Experiment 2 is to contrast this force with interference from the alternative hypothesis by varying $\pi(w|o)$ and by setting $\pi(w|o) \geq 0.5$. This is key, for it helps assess whether interference from the alternative hypothesis may be a source of underreaction to data as described in Figure 3.[30]

[30] We did not include an *IM* treatment due to sample size limitations.

Table 2 described all treatments, identified by the acronym of the true answer N(eutral), I(ntermediate), C(ommon) and interference from irrelevant data, L(ow), M(iddle) and H(igh).

Table 2: Treatments in Experiment 2

Treatment	Distribution	Distribution of Irrelevant Data	Sample Sizes	Elicited Belief
Neutral	50% Orange Words, 50% Orange Numbers	NL: 0% Blue Words	$N = 147$	P(Word | Orange)
		NM: 50% Blue Words	$N = 146$	
		NH: 100% Blue Words	$N = 151$	
Intermediate	55% Orange Words, 45% Orange Numbers	IL: 0% Blue Words	$N = 158$	P(Word | Orange)
		IH: 100% Blue Words	$N = 154$	
Common	70% Orange Words, 30% Orange Numbers	CL: 0% Blue Words	$N = 154$	P(Word | Orange)
		CM: 30% Blue Words	$N = 149$	
		CH: 100% Blue Words	$N = 144$	

Notes: This table describes the treatments in Experiment 2. For all treatments, the L and H sub-treatments consist of 0% and 100% blue words respectively. The *Neutral* and *Common* treatments also have an M sub-treatment, which is 50% blue words for *Neutral* and 30% for *Common.*

4.4 Experiment 2 Results

Panel A of Figure 5 reports, for each treatment, the difference between the average assessment $\hat{\pi}(w|o)$ and the true fraction $\pi(w|o)$ of orange images that are words. Panel B reports the average number of orange words recalled by subjects in each treatment.

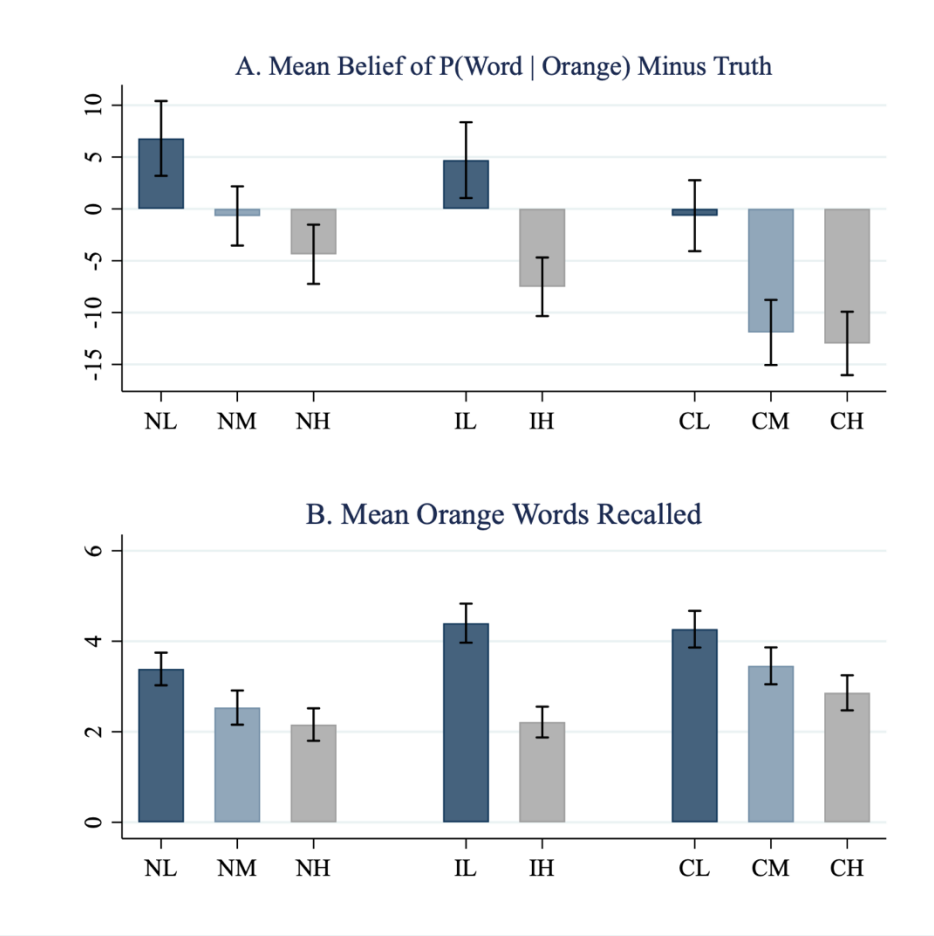

Figure 5: Testing Prediction 4

Notes: Panel A shows the average belief that the randomly drawn image is a word conditional on it being orange minus the true conditional probability. Panel B shows the average number of correctly recalled orange words. In the *L* treatments, all blue images are words. In the *H* Treatments, all blue images are numbers. In the *M* treatment when 70% of orange images are words (CM), 30% of blue images are words. In the *M* treatment when 50% of orange images are words (NM), 50% of blue images are also words. Bands show 95% confidence intervals.

Consistent with our model, stronger interference from irrelevant data (higher $\pi(w|b)$) reduces the assessment $\hat{\pi}(w|o)$ that an orange image is a word, across all treatments ($p < 0.01$ in each case). When normatively irrelevant blue words are more numerous, interference in recall of

orange words is stronger. Recall data in Panel B supports this mechanism: subjects recall fewer correct orange words when $\pi(w|b)$ is higher.[31]

In line with predictions, overestimation of $\hat{\pi}(w|o)$ arises only if orange words are rare enough, namely in the *Neutral* and *Intermediate* treatments. In the *Common* treatments, when $\pi(w|o) = 0.7$, there is no overestimation of $\hat{\pi}(w|o)$ even in the extreme case of $\pi(w|b) = 0$.[32]

In sum, Experiment 2 is consistent with two key predictions of Proposition 4. First, underestimation of a likely hypothesis can be turned into overestimation if the recall of the alternative hypothesis faces strong interference from irrelevant, yet sufficiently similar, experiences. This is evident from the switch from underestimation to overestimation of $\pi(w|o) = 0.55$ as we move from treatment IH to IL (and the consistent drop in the recall or orange words).

Second, if the hypothesis becomes very likely (as in the *Common* treatments), overestimation disappears because interference from the alternative hypothesis becomes very strong. This occurs even if interference from irrelevant data $\pi(w|b)$ is very low, as evident in treatments CL and CM. Treatment CM shows another key prediction of our model: when orange is a weaker signal of the image being a word, beliefs *underreact* to the orange data, $\hat{\pi}(w|o) < \pi(w|o)$, while they are well calibrated when the signal is strong in CL, here $\hat{\pi}(w|o) \approx \pi(w|o)$. In the other treatments, when the data is indicative of a rare hypothesis, beliefs overreact, as

[31] This result, unlike the others in this section, looks different if we focus only on the number of words that participants list in the recall tasks (as opposed to counting the number of *correctly* recalled words). Participants actually list more words as being orange in high interference sub-treatments, though significantly fewer *correct* orange words. We think that this occurs because it is much easier to guess words that may have been orange in treatments in which there are more words overall. This issue does not arise in Experiment 1 (which occurred chronologically after Experiment 2) because there we focus on categories for which it is difficult to incorrectly list a word as being in the wrong category.

[32] These results cannot be explained by the fact that subjects misinterpret our request for P(Word | Orange) as asking for either P(Orange) or P(Orange Word). If so, their answers should not depend on the distribution of blue words. If they interpreted the question as asking for P(Word), the effect should be opposite to what we observe.

predicted by our theory.[33] The balancing of interference from the alternative hypothesis and from irrelevant data accounts for both over and under reaction of beliefs to data. Regularities in selective memory unify different biases in probability judgments.

5. Applications: Similarity and Interference in Economic Decisions

The mechanisms of memory speak to many economic settings. In this section, we discuss some of them, such as saving decisions, the pricing of insurance and Arrow-Debreu securities, and labor market stereotypes. We start with a general setup. A DM evaluates an action a, which yields payoff $u(a)$ today and a state-contingent payoff $u_s(a)$ tomorrow, with the probability of state $s \in \{1, 2\}$ given by $\pi_1(a)$ and $\pi_2(a)$, respectively. The expected utility of action a is:

$$V(a) = u(a) + \sum_{s\in\{1,2\}} u_s(a)\pi_s(a)\,. \tag{10}$$

The action a could be the decision to save, to purchase a security with state-contingent payoffs, or to hire a worker with a particular skill.

Equation (10) highlights a key feature of standard models: the expected utility of an action only depends on its influence on the payoff in each state s and its objective probability. In other words, the payoffs and the objective probability of each state are sufficient statistics for valuation, which is thus invariant to other features of the environment.[34] Similarity in recall breaks down this invariance in two important ways. First, the subjective probability of a state does not just depend on its objective probability, but also on the similarity of experiences associated with each state,

[33] In NH, IH, and CH the data $D =$ "orange" is informative of $H_2 =$ "number". Because in our treatments $\pi(n|o) < 0.5$, here we are in the lower part of Region A, in which the data point to an unlikely hypothesis. Consistent with the model, in these treatments we see overreaction: $\hat{\pi}(n|o)$ is overestimated or equivalently $\hat{\pi}(w|o)$ is underestimated.

[34] This is also true for Prospect Theory and its extensions (e.g. Kőszegi and Rabin 2006), where the subjective decision weights and the reference points depend only on the objective probabilities and the payoffs.

which can be influenced by its description. More homogeneous states are easier to retrieve and receive greater decision weights. Second, the DM's beliefs and actions also depend on the interference from experiences irrelevant to the decision at hand, such as those regarding a counterfactual group or action. In the following applications, we study the implications of these two violations of invariance: our results on savings and Arrow-Debreu security prices highlight the former, while those on social stereotypes and labor market discrimination highlight the latter.

5.1 Similarity, Interference and Savings Decisions

A growing body of work connects under-saving to cognitive mistakes rather than to present bias and impatience (Laibson 1997). Consumers systematically underestimate future expenditures, a phenomenon referred to as the "planning fallacy" (Peetz and Buehler 2009). In particular, they fail to account for idiosyncratic events such as a speeding ticket, a medical need, or a car repair (Sussman and Alter 2012). A *Wall Street Journal* column advises that retirement spending averages $400 more per month than expected because of surprising outlays: "These are bills outside what we normally would expect: the garage door spring and cable that snapped and had to be replaced; the family member who asked for financial help; the X-rays and dentists' fee for a sudden toothache; the small tree in our yard that, it turned out, was dying and needed to be removed; the storm that damaged the screens on our porch; the stone that hit and cracked our windshield; the request from a charity that we felt we needed to honor. The list goes on." (Ruffenach 2022). Failure to account for such events causes under-saving. Augenblick et al. (2022) link such mispredictions of unusual events to savings and spending by farmers in Zambia, which leads to hunger prior to the harvest.

A notable feature of the events described both in the press and academic studies is the extreme diversity of unexpected spending shocks. Our model delivers this phenomenon as a form of systematic forgetting caused by the dissimilarity/heterogeneity of such shocks. Suppose that a is current saving out of normal income Y, so that $u(a) = u(Y - a)$. Future utility is then $u_1(a) = u(Y + a)$ under normal conditions and $u_2(a) = u(Y - L + a)$ under an expenditure shock $L > 0$. $u(.)$ is an increasing and concave Von Neumann-Morgenstern utility function.

The shock hits with exogenous probability $\pi_2 < 1/2$. It can arise from $N > 1$ mutually exclusive causes s_{2i}, $i = 1, \dots, N$, each occurring with probability $\pi_{2,i}$, where $\sum_i \pi_{2i} = \pi_2$. With expected utility in (10), only the total probability π_2 of the shock matters and the DM's optimal choice is simple: savings increase in π_2, $da/d\pi_2 > 0$. As idiosyncratic spending needs become more frequent, the DM transfers more resources to the high marginal utility "shock" state.

Consider now a DM with limited memory. When thinking about saving, the DM estimates the probability of $H_2 =$ "s_2" by retrieving each shock s_{2i} experienced in the past. In his database, s_{2i} is encoded in terms of its cause and the income loss it was associated with. Formally, s_{2i} is a vector of $N + 1$ features. The first feature takes the value 1 because loss L was borne. Feature $i + 1$ takes value 1 because the loss was of type i. All other features are 0. The normal income state s_1 is then a vector of $N + 1$ zeroes. Vectors are encoded with their frequencies $(\pi_{2i}), \pi_1$.

Similarity among experience-vectors shapes retrieval. The self-similarity of an experience is maximal and equal to one, $S(s_1, s_1) = S(s_{2,i}, s_{2,i}) = 1$. Any pair of other experiences, by contrast, differ along two features. The normal state s_1 and any shock state $s_{2,i}$ differ along the occurrence of the loss L and its cause i. Shocks $s_{2,i}$ and $s_{2,j}$ differ on their causes: i and j. Denote by $\Delta \in [0,1]$ the drop in similarity entailed by two dissonant features. We then have $S(s_1, s_{2,i}) = S(s_{2,i}, s_{2,j}) = 1 - \Delta$ for $i = 1, \dots, N$ and $i \neq j$. This in turn implies:

$$S(s_1, s_2) = 1 - \Delta, \tag{11}$$

$$S(s_2, s_2) = 1 - \Delta(1 - C), \tag{12}$$

where $C = \sum_i \left(\frac{\pi_{2,i}}{\pi_2}\right)^2$ is the Herfindhal index of concentration of the aggregate shock state s_2 across the N different causes of expenditure shocks.

Equation (12) embeds the key interference mechanism. The DM's ability to estimate the overall probability of the income loss L depends on how heterogeneous the loss state is. If s_2 is fully concentrated on a single cause, $C = 1$, self-similarity is maximal, $S(s_2, s_2) = 1$, and s_2 is easy to recall. If s_2 is fully dispersed among infinitely many idiosyncratic spending needs, $C \to 0$, then the self-similarity of s_2 is minimal, so experiences of income loss are harder to retrieve.

By plugging (11) and (12) into (5), the total estimated probability of experiencing an expenditure shock is given by:

$$\hat{\pi}_2(\pi_2, C) = \frac{\pi_2(1 - \Delta\pi_2)}{(1 - \pi_2)\left[1 - \frac{\Delta C}{1 - \Delta(1 - C)}\,(1 - \pi_2)\right] + \pi_2(1 - \Delta\pi_2)}. \tag{13}$$

In the absence of similarity-driven distortions, $\Delta = 0$, the DM is well calibrated: $\hat{\pi}_2(\pi_2, C) = \pi_2$. If $\Delta > 0$, his belief is distorted: $\hat{\pi}_2(\pi_2, C) \neq \pi_2$. More specifically, if the causes of income loss are sufficiently heterogeneous and dispersed,

$$C < C^* \equiv \frac{(1 - \Delta)\pi_2}{1 - (1 + \Delta)\pi_2}, \tag{14}$$

the DM underestimates the frequency of the shock, $\hat{\pi}_2(\pi_2, C) < \pi_2$, and overestimates it otherwise.

Given that savings increase in the estimated probability $\hat{\pi}_2(\pi_2, C)$, similarity has important economic implications. If rainy days were due to a single cause, $C = 1$, the DM would overestimate their likelihood and so would over-save. This is consistent with overweighting of unlikely shocks in KT's probability weighting function. But when rainy days come for many

different reasons, C is low, recall of each specific reason faces a lot of interference, which causes forgetting. As a result, the DM underestimates the likelihood of s_2 and under-saves, consistent with the evidence we discussed. Savings decisions no longer satisfy the invariance of Equation (10), and now depend on the similarity of anticipated future expenditures.

Similarity also generates "framing" effects: the DM will save more if specific shocks are described to him, because this frame boosts retrieval (as in Corollary 2). Obermayer et al. (2022) find that such intervention indeed increases savings, and Peetz, Buehler, and Koehler (2014) show that it increases predicted future spending. This mechanism also accounts for the "planning fallacy" (Kahneman and Tversky 1979), in which people systematically underestimate the time required to complete a task. The causes of delay are all different, which hinders their recall.

5.2 Similarity and asset prices

Selective memory has implications for pricing financial assets. Suppose that, rather than saving, the DM can purchase insurance against future income shocks. Formally, the DM chooses the quantity a of Arrow-Debreu securities on shock i to buy, where the security pays off L if a loss in state s_{2i} materializes. Denote the price of such a claim by P_i. If Arrow-Debreu securities are in zero net supply, the rational equilibrium price of insurance against shock i is given by:

$$P_i^r = \mu \cdot \pi_{2,i} \cdot L,$$

where $\mu \equiv u'(Y-L)/u'(Y) > 1$ is the DM's marginal rate of substitution. Furthermore, prices are additive under rationality: the price of buying a broad insurance contract against any income loss of L is equal to the sum of the prices of all Arrow Debreu claims, given by $P^r = \mu \cdot \pi_2 \cdot L$.

Selective memory creates a wedge between these prices. Suppose that all shocks are equally likely, $\pi_{2,i} = \pi_2/N$, and send $N \mapsto \infty$. The shock state is then fully dispersed, $C \mapsto 0$, and the price of broad insurance is given by:

$$P_b = \mu \cdot \hat{\pi}_2(\pi_2, 0) \cdot L = \mu \cdot L \cdot \left(\frac{1 - \Delta\pi_2}{1 - \Delta\pi_2^2}\right)\pi_2, \tag{15}$$

which is less than the rational price P^r for any $\Delta > 0$. Intuitively, the DM fails to retrieve the different shocks insured by the broad contract, and so undervalues that contract.

Consider instead the price of insuring any loss by buying Arrow Debreu claims. The market price of doing so is the price of N identical claims, each one fully concentrated on s_{2i} and paying with probability π_2/N. That is, $P_i = \mu \cdot L \cdot \hat{\pi}_2(\pi_2/N, 1)$, so the total price of all claims is:

$$\lim_{N\to\infty} \mu \cdot L \cdot N \cdot \hat{\pi}_2(\pi_2/N, 1) = \mu \cdot L \cdot \left(\frac{1}{1 - \Delta}\right)\pi_2. \tag{16}$$

In stark contrast with the broad contract, the individual claims are overvalued compared to their rational price $P^r = \mu \cdot L \cdot \pi_2$. This is again due to similarity: as the DM thinks about each specific shock, he focuses on its occurrence, overestimating the insurance pay-out rate.

Similarity causes market prices to be sub-additive, again breaking down the invariance of Equation (10) with respect to fine state descriptions. Similarity explains why people are reluctant to buy broad (e.g., health) insurance but overpay for insuring specific unlikely risks, as documented for extended warranties (Abito and Salant 2018), flight insurance (Eisner and Strotz, 1961), and specific diseases/causes of death (Kunreuther and Pauly 2006, Johnson, Hershey, Meszaros, and Kunreuther 1993). Relatedly, people are more likely to buy insurance after a disaster hits, and gradually cancel the insurance if the policy has not paid out over time (Kunreuther and Pauly 2006). This is in line with the intuition that the possibility of disaster is cued and hence

retrieved only right after its occurrence. More broadly, by generating sub-additive prices, selective memory can have an important impact on asset markets.[35]

5.3 Minority stereotypes and illusory correlation

There is growing interest in economics in understanding social stereotypes, which shape discrimination in labor markets (Neumark 2018), education (Carlana 2019), judicial decisions (Arnold, Dobbie, and C. Yang 2018), and politics (Bonomi, Gennaioli, and G. Tabellini 2021, Bordalo, M. Tabellini, and D. Yang 2020). Our model accounts for the "kernel of truth" model of stereotypes in Bordalo et al. (2016), but also helps explain additional findings from social psychology, which highlight that stereotypes are often directed at minorities (Hilton and von Hippel 1996), and may arise even in the absence of any group differences, as a form of an illusory correlation (Sherman, Hamilton, and Roskos-Ewoldsen 1989).[36]

Before we present the analysis, consider the following example. A board of directors considers a female candidate for a CEO position. Suppose that most CEOs are competent, but also that the vast majority of current CEOs are male. When considering a female candidate, the hypothesis that she is competent suffers strong interference from the large number of male CEOs, who dominate these positions. As a consequence of such interference from irrelevant data, the hypothesis that a female CEO candidate is competent might be underestimated. Interference from very common but irrelevant data becomes a source of an illusory correlations and stereotypes.

[35] For instance, this mechanism may help explain why investors are slower to evaluate news about a company that is "complicated" and consists of many heterogeneous subsidiaries, relative to "pure play" businesses (e.g. Cohen and Lou 2012), or why in a spinoff the parent is valued less than the equity carve out it owns (Lamont and Thaler 2003).
[36] This effect, originally documented in the context of erroneous clinical judgments (Chapman 1967) is robustly produced in experiments, and has been proposed as a mechanism for negative views on minorities as well as for beliefs in non-social settings, e.g. that poor weather is correlated with joint pain (Jena et al 2017).

Formally, an employer decides whether to hire a worker from a minority group G, based on his beliefs about the worker's productivity. In terms of Equation (10), hiring the worker ($a = 1$) yields high utility $u_1(a) = \theta_H a$ if the worker is productive, which occurs with probability $\pi_{H,G}$, and low utility $u_2(a) = -\theta_L a$ if the worker is unproductive (with probability $\pi_{L,G} = 1 - \pi_{H,G}$). For simplicity, we assume no utility cost in hiring ($u(a) = 0$). A rational DM hires the worker if the probability he is unproductive is low enough, $\pi_{L,G} < \pi^* \equiv \theta_H/(\theta_H + \theta_L)$.

A DM with selective memory forms belief $\hat{\pi}_{L,G}$ by sampling his past experiences. The memory database encodes two features: whether or not a worker is productive, H or L, and his social group, G or $\overline{G}$. Experiences that share only one feature have similarity $1 - \Delta$, while those differing in both features have similarity $1 - 2\Delta$. The extent of interference depends on the prevalence of the groups, denoted by p_G and $p_{\overline{G}} = 1 - p_G$ respectively, as well as of the low type in $\overline{G}$, denoted by $\pi_{L,\overline{G}}$. Using (5), the estimated odds that a G worker has low productivity is:

$$\frac{\hat{\pi}_{L,G}}{\hat{\pi}_{H,G}} = \frac{\pi_{L,G}}{\pi_{H,G}} \cdot \frac{p_G\Delta\left(1 + \pi_{H,G} - \pi_{H,\overline{G}}\right) + \Delta\pi_{H,\overline{G}} + (1 - 2\Delta)}{p_G\Delta\left(1 + \pi_{L,G} - \pi_{L,\overline{G}}\right) + \Delta\pi_{L,\overline{G}} + (1 - 2\Delta)}. \tag{17}$$

If $\Delta > 0$, the DM overestimates the probability that the worker from G is a low type if and only if:

$$\pi_{L,G} < \varphi \equiv \frac{\frac{\pi_{LG}}{\pi_{L\overline{G}}}}{p_G\frac{\pi_{LG}}{\pi_{L\overline{G}}} + p_{\overline{G}}}. \tag{18}$$

The DM has a negative stereotype of the worker from G if the share of experiences with low types from this group ($\pi_{L,G}$) is smaller than a threshold φ. In line with Proposition 4, low frequency events tend to be overestimated. Crucially, this threshold depends on two features of the database. First, it increases in the likelihood ratio of low types in G relative to $\bar{G}$, $\pi_{LG}/\pi_{L\overline{G}}$. The likelihood ratio is high when most members of the majority group $\bar{G}$ are high types ($\pi_{L\overline{G}}$ is low),

so that they strongly interfere with the recall of high types in the minority group G. In turn, this causes an overestimation of low types in G. Stereotypes are an example of the second type of violation of the invariance in Equation (10): the retrieval of irrelevant experiences in $\overline{G}$ intrude and distort assessments.

Interference provides a memory foundation for the stereotypes model of BCGS (2016), which also relies on the likelihood ratio, but also generates new predictions. In particular, it predicts that stereotypes should be stronger for minority because the strength of interference from $\overline{G}$ is especially high when the majority dominates the database, i.e. when $1 - p_G$ is high. This implies that, provided $\pi_{L,G} \geq \pi_{L,\overline{G}}$, the low stereotype is exacerbated for minorities: Equation (18) is easier to meet if p_G is low.

This effect can produce minority stereotypes even if different types are equally frequent in both groups, a phenomenon known as illusory correlation (Sherman, Hamilton, and Roskos-Ewoldsen 1989). To see this, set $\pi_{L,G} = \pi_{L,\overline{G}} = \pi_L$ in Equation (17) to obtain:

$$\frac{\hat{\pi}_{L,G}}{\hat{\pi}_{H,G}} = \frac{\pi_L}{\pi_H} \cdot \frac{p_G\Delta + 1 - \Delta - \Delta\pi_L}{p_G\Delta + 1 - \Delta - \Delta(1 - \pi_L)}, \tag{19}$$

When low types are rare, $\pi_L < 1/2$, their frequency is overestimated if and only if the group is a minority, $p_G < 1/2$. Notably, the stereotype emerges *even though the share of low types in the two groups is the same*. Recall of high types in G is inhibited by the many high types from $\overline{G}$ that flood the DM's memory database, while the reverse interference from G to $\bar{G}$ is far weaker.

6. Conclusion

We have presented a model of memory-based probability judgments, with two main ingredients: i) databases of experiences, and ii) cues that trigger selective recall of these

experiences. Recall is driven by similarity of the experiences to the cues, which include both hypotheses and data. Similarity helps retrieve relevant experiences, but also invites interference from experiences inconsistent with the hypothesis at hand (but similar to it). The new insight is that a hypothesis is underestimated when, compared to its alternative, it is more vulnerable to interference because it is more heterogeneous, more likely, or more similar to irrelevant data.

This notion that probability estimates are shaped by content (as captured by feature similarity) and not just by objective frequency accounts for and reconciles a wide range of seemingly inconsistent experimental and field evidence, including availability and representativeness heuristics proposed by Kahneman and Tversky (1974), overestimation of the probabilities of unlikely hypotheses, conjunction and disjunction fallacies in experimental data, as well as under and overreaction to information. We tested several novel predictions of the model using an experimental design in which we control both the memory database and the cues subjects receive, and found strong supportive evidence. Finally, we showed how memory-based beliefs shed light on several economic applications, linking under-saving and sub-additivity of prices to failure to forecast heterogeneous states of the world under a broad cue, and minority stereotypes to interference from the larger majority group.

The analysis in this paper opens the gates for many research directions, and in conclusion we list three we find particularly promising. First, probability judgments can pertain to events not yet experienced by the decision maker, such as forecasts of the future, or to events that are described in terms of statistics or data generating processes (Benjamin 2019). Memory plausibly plays a central role in these settings as well. With respect to forecasts, a significant literature in psychology shows that the mental simulation of future events is intimately linked to memory processes (Dougherty et al. 1997; Brown et al. 2000). People combine past experiences with

simulated ones (Kahneman and Miller 1986; Schachter et al. 2007; Biderman et al 2020), with the ease of simulation also driven by perceived similarity (Woltz and Gardner 2015). In this way, memory shapes forecasts. Bordalo et al (2022) incorporate simulation into the model presented here and apply it to studying beliefs about Covid, a novel threat. The model explains strong regularities of such beliefs, including the fact that, through interference, experiencing non health adversities leads to less pessimism about Covid lethality for the general population.

In addition, individuals often have both statistical and experiential information, such as in the literature on the description-experience gap in risky choice (Hertwig and Erev 2009). This research suggests an interaction between the two sources of information in generating beliefs, where statistical information may also act as a cue for retrieving semantic content from memory.

Expanding the model may also lead to new predictions. One important direction is to better understand the drivers of retrieval, here summarized by a similarity function. Different people may interpret the same cue differently, depending in part on differences in their experiences, on their perceptions of similarity or attention to features of the stimulus, or on chance. The attention channel can be important. In an experiment with U.S. federal judges by Clancy et al. (1981), judges adjudicated a set of hypothetical criminal cases with multiple attributes. The authors found that different judges attended to different attributes of the case and proposed radically different sentences. Such heterogeneous responses may naturally occur if a decision maker's perceived similarity depends on the range of past experiences, or if these experiences influence the mental model that the decision maker uses (Schwartzstein 2014).

Another theoretical extension concerns learning and its distortions. For example, in our approach signals about an event prime recall of previous experiences of the event itself, which may create a form of confirmation bias (Nickerson 1998).

Finally, our analysis focuses on the role of memory in probability estimates, but the applications of cued recall based on similarity to belief formation are much broader. The principles we described in this paper can be applied to many problems, including consumer choice, advertising, persuasion, political positioning, product branding, and many others.

REFERENCES

Abito, Jose Miguel, and Yuval Salant. "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies* 86 (2019), 2285-2318.

Anderson, Michael, and Barbara Spellman. "On the Status of Inhibitory Mechanisms in Cognition: Memory Retrieval as a Model Case," *Psychological Review*, 102 (1995), 68-100.

Anderson, John, and Lynne Reder. "The Fan Effect: New Results and New Theories," *Journal of Experimental Psychology: General,* 128 (1999), 186-197.

Arnold, David, Dobbie, Will, and Crystal Yang. "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133 (2018), 1885-1932.

Augenblick, Ned, Kelsey Jack, Supreet Kaur, Felix Masiye, and Nicholas Swanson. "Budget Neglect in Consumption Smoothing: A Field Experiment on Seasonal Hunger" Working paper (2022).

Azeredo da Silveira, Rava, Yeji Sung, and Michael Woodford. "Optimally Imprecise Memory and Biased Forecasts," w28075. National Bureau of Economic Research, (2020).

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua Teitelbaum. "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103 (2013), 2499-2529.

Benjamin, Daniel. "Errors in Probabilistic Reasoning and Judgment Biases," *Handbook of Behavioral Economics: Applications and Foundations* 1 (2019), 69-186.

Biderman, N., Bakkour, A., and Daphna Shoham. "What are memories for? The hippocampus bridges past experience with future decisions," *Trends in Cognitive Sciences*, *24*(7) (2020), 542-556.

Billot, Antoine, Itzhak Gilboa, Dov Samet, and David Schmeidler. "Probabilities as Similarity-Weighted Frequencies," *Econometrica,* 73 (2005)*,* 1125-1136.

Bonomi, Giampaolo, Nicola Gennaioli, and Guido Tabellini. "Identity, Beliefs, and Political Conflict," *Quarterly Journal of Economics* 136 (2021), 2371-2411.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer. "Memory and Representativeness," *Psychological Review*, 128 (2020), 71-85.

Bordalo, Pedro, Giovani Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. "Imagining the Future: Memory, Simulation, and Beliefs about Covid," Working paper (2022).

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. "Stereotypes," *Quarterly Journal of Economics*, 131 (2016), 1753-1794.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. "Overreaction in Macroeconomic Expectations," *American Economic Review*, 110(2020), 2748-2782.

Bordalo, Pedro, Nicola Gennaioli, Rafael LaPorta, and Andrei Shleifer. "Diagnostic Expectations and Stock Returns," *Journal of Finance*, 74(2019), 2839-2874.

Bordalo, Pedro, Nicola Gennaioli, Rafael LaPorta, and Andrei Shleifer. "Belief Overreaction and Stock Market Puzzles," Working paper (2022).

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Salience Theory of Choice under Risk," *Quarterly Journal of Economics*, 127 (2012), 1243-1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Diagnostic Expectations and Credit Cycles," *Journal of Finance*, 73(2018), 199-227.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 135(2020), 1399-1442.

Bordalo, Pedro, Marco Tabellini, and David Yang. Issue salience and political stereotypes. NBER w27194 (2020).

Bouchaud, Jean-Philippe, Philipp Krueger, Augustin Landier, and David Thesmar. "Sticky Expectations and the Profitability Anomaly," *Journal of Finance*, 74(2019), 639-674.

Brown, Norman, Lori Buchanan, and Roberto Cabeza. "Estimating the Frequency of Nonevents: the Role of Recollection Failure in False Recognition," *Psychonomic Bulletin and Review*, *7*(2000), 684-691.

Carlana, Michela. "Implicit Stereotypes: Evidence from Teachers' Gender Bias," *Quarterly Journal of Economics*, 134(2019), 1163-1224.

Chan, Louis, Narasimhan Jegadeesh, and Josef Lakonishok. "Momentum Strategies," *Journal of Finance*, 51(1996), 1681-1713.

Chapman, Loren. "Illusory Correlation in Observational Report," *Journal of Verbal Learning and Verbal Behavior*, 6 (1967), 151-155.

Chiappori, Pierre-Andre, Bernard Salanié, François Salanié, and Amit Gandhi. "From Aggregate Betting Data to Individual Risk Preferences," *Econometrica*, 87 (2019), 1-36.

Clancy, Kevin, John Bartolomeo, David Richardson, and Charles Wellford. "Sentence Decision-Making: the Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity," *Journal of Criminal Law and Criminology*, 72 (1981), 524-554.

Coibion, Olivier, and Yuriy Gorodnichenko. "What Can Survey Forecasts Tell Us About Information Rigidities?" *Journal of Political Economy*, 120 (2012), 116-159.

Danz, David, Lise Vesterlund, and Alistair Wilson. "Belief Elicitation: Limiting Truth Telling with Information on Incentives," w27327. National Bureau of Economic Research (2020).

Dasgupta, Ishita, Eric Schulz, Joshua Tenenbaum, and Sam Gershman. "A Theory of Learning to Infer," *Psychological Review*, 127 (2020), 412-441.

Dasgupta, Ishita, and Sam Gershman. "Memory as a Computational Resource," *Trends in Cognitive Sciences.* 25 (2021), 240-251.

Deese, James. "Influence of Inter-Item Associative Strength Upon Immediate Free Recall," *Psychological Reports*, 5 (1959), 305-312.

Dougherty, Michael, Charles Gettys, and Rickey Thomas. "The Role of Mental Simulation in Judgments of Likelihood," *Organizational Behavior and Human Decision Processes*, 70 (1997), 135-148.

Dougherty, Michael, Charles Gettys, and Eve Ogden. "MINERVA-DM: a Memory Processes Model for Judgments of Likelihood," *Psychological Review*, 106 (1999), 180-209.

Edwards, Ward. "Conservatism in human information processing." In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 359-369). Cambridge: Cambridge University Press (1968).

Eisner, Robert, and Robert Strotz. "Flight Insurance and the Theory of Choice," *Journal of Political Economy*, 69 (1961), 355-368.

Enke, Ben, and Thomas Graeber. "Cognitive Uncertainty," w26518. National Bureau of Economic Research (2019).

Enke, Ben, Frederik Schwerter, and Florian Zimmermann. "Associative Memory and Belief Formation," w26664. National Bureau of Economic Research (2020).

Hertwig, Ralph, and Ido Erev. "The Description–Experience Gap in Risky Choice," *Trends in Cognitive Sciences*, 13 (2009), 517-523.

Fischhoff, Baruch, Paul Slovic, and Sarah Lichtenstein. "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation," *Journal of Experimental Psychology: Human Perception and Performance*, 4 (1978), 330-344.

Frydman, Cary, and Lawrence Jin. "Efficient Coding and Risky Choice," *Quarterly Journal of Economics,* 137 (2020), 161-213.

Gabaix, Xavier. "Behavioral Inattention," *Handbook of Behavioral Economics: Applications and Foundations*, 1 (2019), 261-343.

Gennaioli, Nicola, and Andrei Shleifer. "What Comes to Mind," *Quarterly Journal of Economics*, 125 (2010), 1399-1433.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. "Neglected Risks, Financial Innovation, and Financial Fragility," *Journal of Financial Economics*, 104 (2012), 452-468.

Grether, David. "Bayes Rule as a Descriptive Model: the Representativeness Heuristic," *Quarterly Journal of Economics*, 95 (1980), 537-557.

Griffin, Dale, and Amos Tversky. "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24 (1992), 411-435.

Hilton, James, and William Von Hippel. "Stereotypes," *Annual review of psychology* 47 (1996), 237-271.

Jena, Anupam, Andrew Olenski, David Molitor, and Nolan Miller. "Association Between Rainfall and Diagnoses of Joint or Back Pain: Retrospective Claims Analysis," *British Medical Journal*, 359 (2017).

Jenkins, John, and Karl Dallenbach. "Obliviscence During Sleep and Waking," *American Journal of Psychology*, 35 (1924), 605-612.

Johnson, Eric, Gerald Häubl, and Anat Keinan. "Aspects of Endowment: a Query Theory of Value Construction," *Journal of experimental psychology: Learning, memory, and cognition*, 33 (2007), 461.

Johnson, Eric, John Hershey, Jacqueline Meszaros, and Howard Kunreuther. "Framing, Probability Distortions, and Insurance Decisions," *Journal of risk and uncertainty,* 7 (1993), 35-51.

Kahana, Michael. *Foundations of human memory*. Oxford University Press USA (2012).

Kahneman, Daniel, and Shane Fredrick. "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," *Heuristics and biases: The psychology of intuitive judgment,* 49 (2002), 81.

Kahneman, Daniel, Barbara Fredrickson, Charles Schreiber, and Donald Redelmeier. "When More Pain is Preferred to Less: Adding a Better End," *Psychological Science*, 4 (1993), 401-405.

Kahneman, Daniel, and Dale Miller. “Norm Theory: Comparing Reality to its Alternatives,” *Psychological Review*, 93 (1986), 136-153.

Kahneman, Daniel, Olivier Sibony, and Cass Sunstein. *Noise: a flaw in human judgment*. Little, Brown (2021).

Kahneman, Daniel, and Amos Tversky. “On the Psychology of Prediction,” *Psychological Review*, 80 (1973), 237-251.

Kahneman, Daniel, and Amos Tversky. “Prospect Theory: an Analysis of Decision under Risk,” *Econometrica*, 47 (1979), 263-292.

Keppel, Geoffrey. “Verbal Learning and Memory,” *Annual Review of Psychology*, 19 (1968), 169-202.

Khaw, Mel, Ziang Li, and Michael Woodford. “Cognitive Imprecision and Small-Stakes Risk Aversion,” *Review of Economic Studies*, 88 (2020), 1979-2013.

Kőszegi, Botond, and Matthew Rabin. “A Model of Reference-Dependent Preferences,” *Quarterly Journal of Economics*, 121 (2006), 1133-1165.

Kunreuther, Howard, and Mark Pauly. *Insurance decision-making and market behavior*. now publishers Inc (2006).

Kwon, Spencer, and Johnny Tang. “Reactions to News and Reasoning by Exemplars,” *Available at SSRN* (2020).

Laibson, David. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, 112 (1997), 443-478.

Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. “Judged Frequency of Lethal Events,” *Journal of Experimental Psychology: Human Learning and Memory*, 4 (1978), 551-578.

Lohnas, Lynn, Sean Polyn, and Michael Kahana “Expanding the Scope of Memory Search: Modeling Intralist and Interlist Effects in Free Recall,” *Psychological Review*, 122 (2015), 337-363.

Malmendier, Ulrike, and Stefan Nagel. “Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?” *Quarterly Journal of Economics*, 126 (2011), 373-416.

McGeoch, John. “Forgetting and the Law of Disuse,” *Psychological Review*, 39 (1932), 352-370.

Mullainathan, Sendhil. “A Memory-Based Model of Bounded Rationality,” *Quarterly Journal of Economics*, 117 (2002), 735-774.

Mullen, Brian, Jennifer Atkins, Debbie Champion, Cecelia Edwards, Dana Hardy, John Story, and Mary Vanderklok. "The False Consensus Effect: A Meta-analysis of 115 Hypothesis Tests," *Journal of Experimental Social Psychology*, 21 (1985), 262–283.

Neumark, David. "Experimental Research on Labor Market Discrimination," *Journal of Economic Literature* 56 (2018), 799-866.

Nickerson, Raymond. (1998). "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, 2 (1998), 175-220.

Nosofsky, Robert. "Similarity Scaling and Cognitive Process Models," *Annual review of Psychology*, 43 (1992), 25-53.

Pantelis, Peter, Marieke Van Vugt, Robert Sekuler, Hugh Wilson, and Michael Kahana. "Why Are Some People's Names Easier to Learn Than Others? The Effects of Face Similarity on Memory for Face-Name Associations," *Memory and Cognition*, 36 (2008), 1182-1195.

Peetz, Johanna, and Roger Buehler. "Is There a Budget Fallacy? The Role of Savings Goals in the Prediction of Personal Spending." *Personality and Social Psychology Bulletin* 35, no. 12 (2009), 1579-1591.

Roediger, Henry, and Kathleen McDermott. "Creating False Memories: Remembering Words Not Presented in Lists," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21 (1995), 803-814.

Roese, Neal, and Kathleen Vohs. (2012). "Hindsight Bias," *Perspectives on Psychological Science*, 7 (2012), 411-426.

Ruffenach, Glenn. "How Much More Will You Spend in Retirement Than Expected? My Rule: $400 a Month" *Wall Street Journal* 31/3/2022, Available at: https://www.wsj.com/articles/how-much-more-will-you-spend-in-retirement-than-expected-11648685983 (Accessed: 4/3/22)

Sanborn, Adam, and Nick Chater. "Bayesian Brains Without Probabilities," *Trends in Cognitive Sciences*, 20 (2016), 883-893.

Schacter, Daniel, Donna Addis, and Randy Buckner. "Remembering the Past to Imagine the Future: The Prospective Brain," *Nature Reviews Neuroscience,* 8 (2007), 657-661.

Schacter, Daniel, Donna Addis, Demis Hassabis, Victoria Martin, Nathan Spreng, and Karl Szpunar. "The Future of Memory: Remembering, Imagining, and the Brain," *Neuron*, 76 (2012), 677-94.

Schwartzstein, Joshua. "Selective Attention and Learning," *Journal of the European Economic Association*, 12 (2014), 1423-1452.

Sherman, Steven, David Hamilton, and David Roskos-Ewoldsen. "Attenuation of Illusory Correlation," *Personality and Social Psychology Bulletin* 15 (1989), 559-571.

Shiffrin, Richard. "Memory Search," In *Models of Human Memory* (1970), 375-447.

Slamecka, Norman. "An Examination of Trace Storage in Free Recall," *Journal of Experimental Psychology*, 76 (1968), 504-513.

Sloman, Steven, Yuval Rottenstreich, Edward Wisniewski, Constantinos Hadjichristidis, and Craig Fox. "Typical versus Atypical Unpacking and Superadditive Probability Judgment," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30 (2004), 573-582.

Sims, Christopher. "Implications of Rational Inattention," *Journal of Monetary Economics*, 50 (2003), 665-690.

Sussman, Abigail, and Adam Alter. "The Exception is the Rule: Underestimating and Uverspending on Exceptional Expenses," *Journal of Consumer Research*, 39 (2012), 800-814.

Sydnor, Justin. "(Over) Insuring Modest Risks," *American Economic Journal: Applied Economics*, 2 (2010), 177-199.

Tenenbaum, Joshua, and Thomas Griffiths. "Generalization, Similarity, and Bayesian Inference," *Behavioral and Brain Sciences*, 24 (2001), 629-640.

Tversky, Amos. "Features of Similarity," *Psychological Review*, 84 (1977), 327-352.

Tversky, Amos, and Daniel Kahneman. "Availability: a Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, 5 (1977), 207-232.

Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185 (1974), 1124-1131.

Tversky, Amos, and Daniel Kahneman. "Extensional versus Intuitive Reasoning: the Conjunction Fallacy in Probability Judgment," *Psychological Review*, 90 (1983), 293-315.

Tversky, Amos, and Koehler, Derek. "Support Theory: a Nonextensional Representation of Subjective Probability," *Psychological Review*, 101 (1994), 547-567.

Underwood, Benton. "Interference and Forgetting," *Psychological Review*, 64 (1957), 49-60.

Wachter, Jessica, and Michael Kahana. "A Retrieved-Context Theory of Financial Decisions," w26200. National Bureau of Economic Research (2019).

Whitely, Paul. “The Dependence of Learning and Recall upon Prior Intellectual Activities,” *Journal of Experimental Psychology*, 10 (1927), 489-508.

Woltz, Dan, and Michael Gardner. “Semantic Priming Increases Word Frequency Judgments: Evidence for the Role of Memory Strength in Frequency Estimation,” *Acta Psychologica*, 160 (2015), 152-160.

APPENDIX A

In Appendix A0, we present a summary of the results in the main text. In Appendix A1, we present proofs for the results in the main text. In Appendix A2 we present further results, including i) a generalization of the model to multiple hypotheses, and ii) a proof of the equivalence between the definition of under or overreaction in Section 3.3 and the Coibion and Gorodnichenko test (Coibion and Gorodnichenko 2015).

A0: Summary of results

Bias	Assumptions on S
Smearing (Proposition 2): Small probabilities are overestimated, and large probabilities are underestimated.	Self-similarity is higher than cross-similarity: $S(H_i, H_i) \geq S(H_i, H_j)$
Neglect of non-cued rare events (Corollary 1): An unlikely sub-hypothesis ($H_{21} \subset H_2$) that is not explicitly cued is underestimated.	The similarity of the sub-hypothesis to the broader hypothesis is lower than the self-similarity of the broader hypothesis ("the sub-hypothesis is atypical"): $S(H_{21}, H_2) < S(H_2, H_2)$
Disjunction effect (Proposition 3): Partitioning a hypothesis (H_2) into more self-similar hypotheses ($H_{21} \cup H_{22} = H_2$) increases the estimation of H_2.	Self-similarity of the subhypotheses are higher than their cross-similarities (the subhypotheses form a more homogeneous group): $S(H_{21}, H_{21}) > S(H_{21}, H_{22})$
Underestimation of heterogeneous hypothesis (Corollary 2): Hypotheses with lower self-similarity are likelier to be underestimated, even if the hypothesis is rare.	Holds generally.
Under-and-overreaction (Proposition 4, Corollary 3): If D raises (lowers) the probability of H_i, the agent overreacts (underreacts) if $\pi(H_i \mid \overline{D})$ is sufficiently low, and underreacts (overreacts) if $\pi(H_i \mid D)$ is sufficiently high.	There is more interference from irrelevant experiences that are in the same hypothesis, than irrelevant experiences in a different hypothesis: $S(H_i \cap D,\ H_i \cap \overline{D}) > S\left(H_i \cap D, H_j \cap \overline{D}\right),$ For $i, j = 1, 2,\ i \neq j$.

Table A1: Summary of Main Results

Notes: Table specifies the main biases predicted by the model, and the required assumptions regarding the similarity function.

Table A1 summarizes the main biases predicted by our model, and the required assumptions on the similarity function. In general, the assumptions are satisfied when each hypothesis and sub-hypotheses form a relatively more homogeneous subset relative to the broader set of experiences. The feature-based approach taken in Section 5 (closely related to Tversky 1977) naturally satisfies such conditions.

A1: Proofs

Proposition 1. Let $r(H_i)$ be the recall fluency of hypothesis $i = 1,2$. The sampling process implies that, with samples of size T, the number of successes N_i in recalling hypothesis $i = 1,2$ follows a binomial $Bin\big(T, r(H_i)\big)$. Then, by the central limit theorem, one can take the following asymptotic approximations $\frac{N_i - Tr(H_i)}{\sqrt{T}} = z_i \sim N\left(0, r(H_i)\big(1 - r(H_i)\big)\right)$. Then, noting that the odds in favour of H_1 are equal to N_1/N_2 we see that these follow:

$$\frac{N_1}{N_2} \sim \frac{r(H_1) + \frac{z_1}{\sqrt{T}}}{r(H_2) + \frac{z_2}{\sqrt{T}}} \approx \frac{r(H_1)}{r(H_2)} - \frac{r(H_1)}{r(H_2)^2} \cdot \sqrt{\frac{1}{T}} z_2 + \frac{1}{r(H_2)} \cdot \sqrt{\frac{1}{T}} z_1 + O\left(\frac{1}{T}\right).$$

If T is large enough, given the distribution of z_i, the following normal approximation holds:

$$\frac{N_1}{N_2} \to N\left(\frac{r(H_1)}{r(H_2)}, \frac{1}{T} \cdot \left(\frac{r(H_1)^2}{r(H_2)^4} \cdot r(H_2)\big(1 - r(H_2)\big) + \frac{1}{r(H_2)^2} \cdot r(H_1)\big(1 - r(H_1)\big)\right)\right)$$

which can be written in the form of Proposition 1. ■

Proposition 2. First, we prove that $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}$ is monotonically increasing in $\pi(H_1)$: note that $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)} = \frac{\pi(H_1)}{\pi(H_2)} \cdot \frac{\pi(H_2)+\frac{S(H_1,H_2)}{S(H_2,H_2)}\pi(H_1)}{\pi(H_1)+\frac{S(H_1,H_2)}{S(H_1,H_1)}\pi(H_2)}$, with $\pi(H_2) = 1 - \pi(H_1)$. Taking the derivative of $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}$ with respect to $\pi(H_1)$ yields:

$$\frac{d\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}}{d\pi(H_1)} \propto \frac{S(H_1,H_2)}{S(H_2,H_2)}\pi(H_1)^2 + \frac{S(H_1,\ H_2)}{S(H_1,H_1)} \cdot (1-\pi(H_1)) \cdot \left(\left(2 \cdot \frac{S(H_1,H_2)}{S(H_2,H_2)} - 1\right) \cdot \pi(H_1) + 1\right)$$

The first term is clearly positive. As $0 \leq \pi(H_1) \leq 1$, it suffices to show:

$$\left(2 \cdot \frac{S(H_1,H_2)}{S(H_2,H_2)} - 1\right) \cdot \pi(H_1) + 1 \geq 0$$

Which holds because $\frac{S(H_1,H_2)}{S(H_2,H_2)} \geq 0$. Thus $\left(2 \cdot \frac{S(H_1,H_2)}{S(H_2,H_2)} - 1\right) \cdot \pi(H_1) + 1 \geq -\pi(H_1) + 1 \geq 0$.

Second, the above expression shows that for $0 < \pi(H_1) < 1$ and $S(H_1, H_2) > 0$, we have that $\frac{d\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}}{d\pi(H_1)} > 0$, and thus we have strict monotonicity. Furthermore, for $\pi(H_1) \neq 0, 1$ and $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)} = \frac{\pi(H_1)}{\pi(H_2)}$, we have:

$$\pi(H_1) + \frac{S(H_1,\ H_2)}{S(H_1,H_1)}\ \pi(H_2) = \pi(H_2) + \frac{S(H_1,H_2)}{S(H_2,H_2)}\ \pi(H_1) \Leftrightarrow \left(1 - \frac{S(H_1,H_2)}{S(H_2,H_2)}\right)\pi(H_1)$$

$$= \left(1 - \frac{S(H_1,H_2)}{S(H_1,H_1)}\right)\pi(H_2),$$

which implicitly defines π^*. Finally, note that:

$$\hat{\pi}(H_1) > \pi(H_1) \Leftrightarrow \frac{\pi(H_2)+\frac{S(H_1,H_2)}{S(H_2,H_2)}\pi(H_1)}{\pi(H_1)+\frac{S(H_1,H_2)}{S(H_1,H_1)}\pi(H_2)} > 1 \Leftrightarrow \left(1 - \frac{S(H_1,H_2)}{S(H_2,H_2)}\right)\pi(H_1) < \left(1 - \frac{S(H_1,H_2)}{S(H_1,H_1)}\right)\pi(H_2).$$

Assuming that $S(H_1, H_2) < S(H_2, H_2)$, the above is satisfied iff $\pi(H_1) < \pi^*$, as desired. ■

Corollary 1. Let $r(A; H_2)$ denote the probability of sampling event A when thinking about H_2. Then, the share of successful recall events for H_2 that fall in subset $H_{21} \subset H_2$ are:

$$\frac{r(H_{21}; H_2)}{r(H_2)} = \frac{S(H_{21}, H_2)\pi(H_{21})}{S(H_2, H_2)\pi(H_2)}.$$

H_2' is undersampled relative to its true frequency if and only if $S(H_2', H_2) < S(H_2, H_2)$.

Given that: $S(H_2, H_2) = S(H_{21}, H_2) \cdot \pi(H_{21}|H_2) + S(H_2 \setminus H_{21}, H_2)\pi \cdot (H_2 \setminus H_{21}|H_2)$, this condition can be written as $S(H_{21}, H_2) < S(H_2 \setminus H_{21}, H_2)$. In turn, note that:

$$S(H_{21}, H_2) = S(H_{21}, H_2')\pi(H_{21}|H_2) + S(H_{21}, H_2 \setminus H_{21})\pi(H_2 \setminus H_{21}|H_2)$$

$$S(H_2 \setminus H_{21}, H_2) = S(H_{21}, H_2 \setminus H_2')\pi(H_{21}|H_2) + S(H_2 \setminus H_{21}, H_2 \setminus H_{21})\pi(H_2 \setminus H_{21}|H_2)$$

Thus, since $S(A, B) = S(B, A)$, we have that $S(H_{21}, H_2) < S(H_2 \setminus H_{21}, H_2)$ if and only if:

$$[S(H_{21}, H_{21}) - S(H_{21}, H_{22})] \cdot \pi(H_{21}|H_2) < [S(H_{22}, H_{22}) - S(H_{21}, H_{22})] \cdot \pi(H_{22}|H_2)$$

If the events are equally self-similar, $S(H_{21}, H_{21}) = S(H_2 \setminus H_{21}, H_2 \setminus H_{21})$, given that self similarity is larger than cross similarity, the condition holds if and only if $\pi(H_{21}|H_2) < \pi(H_2 \setminus H_{21}|H_2)$. ■

Corollary 2. The proof follows from the inspection of the expression:

$$\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)} = \frac{\pi(H_1)}{\pi(H_2)} \cdot \frac{\pi(H_2) + \frac{S(H_1, H_2)}{S(H_2, H_2)}\ \pi(H_1)}{\pi(H_1) + \frac{S(H_1,\ H_2)}{S(H_1, H_1)}\ \pi(H_2)} \blacksquare$$

Proposition 3. Given two subsets H_{21} and H_{22}, with three hypotheses (H_1, H_{21}, H_{22}) the recall fluency of H_1, $r(H_1)$, is the same as when there are only hypotheses H_1 and H_2. We then have:

$$r(H_{2i}) = \frac{S(H_{2i}, H_{2i})\pi(H_{2i})}{S(H_{2i}, H_{2i})\pi(H_{2i}) + S\big(H_{2i}, H_{2j}\big)\pi\big(H_{2j}\big) + S(H_{2i}, H_1)\pi(H_1)}.$$

Denote $S(H_{2i}, H_{2i}) = S^*$ and $S(H_{2i}, H_1) = S_*$ which are by assumption independent from $i = 1,2$. We can then write:

$$r(H_{2i}) = \frac{1}{2}\frac{S^*\pi(H_2)}{\frac{\left[S^* + S\big(H_{2i}, H_{2j}\big)\right]}{2}\pi(H_2) + S_*\pi(H_1)}.$$

Given the symmetry of similarity, $r(H_{21}) = r(H_{22})$. As a result, the probability of successes in sampling H_2, namely the sum of successes in sampling H_{21} and H_{22} follows a binomial distribution $Bin\big(2T, r(H_{21})\big)$. As a result, when the hypotheses is split into the two subsets, the average number of successes in recalling H_2 per attempt T is:

$$r'(H_2) = \frac{S^*\pi(H_2)}{\frac{\left[S^* + S\big(H_{2i}, H_{2j}\big)\right]}{2}\pi(H_2) + S_*\pi(H_1)}.$$

When the hypothesis is not split, we have:

$$r(H_2) = \frac{S(H_2, H_2)\pi(H_2)}{S(H_2, H_2)\pi(H_2) + S(H_2, H_1)\pi(H_1)},$$

which, given the assumption, $S(H_{21}, H_1) = S(H_{22}, H_1) = S_*$ is equal to:

$$r(H_2) = \frac{S(H_2, H_2)\pi(H_2)}{S(H_2, H_2)\pi(H_2) + S_*\pi(H_1)}.$$

Because the two subsets are equally likely, it is immediate to find that $\frac{\left[S^* + S\big(H_{2i}, H_{2j}\big)\right]}{2} = S(H_2, H_2)$. It then immediately follows that $r'(H_2) > r(H_2)$ if and only if $S^* > S(H_2, H_2)$, which is equivalent to $S(H_{2i}, H_{2i}) > S\big(H_{2i}, H_{2j}\big)$. ■

Proposition 4. We prove the result of the under and overestimation of conditional beliefs for general similarity specification. The individual is assessing $\pi(H_1|D)$. For notational simplicity,

we shall denote the four sub-populations as follows: $H_1D, H_2D, H_1\overline{D}, H_2\overline{D}$, and we shall use as a loose-hand $S(A,B)$, where $A, B \in \{H_1D, H_2D, H_1\overline{D}, H_2\overline{D}\}$, the similarity between any two given sub-populations.

We assume for simplicity that the self-similarity of all subgroups is equal to 1, and $\pi(D) = \pi(\overline{D})$ Furthermore, we assume the following inequalities regarding the similarity between subgroups:

$$S(H_iD,\ H_i\overline{D}) > S(H_iD, H_{-i}\overline{D}),$$

Where $i = 1, 2$ and $-i = 2, 1$. To give an intuition behind this assumption, let us use the example in the paper $H_1 =$ "accident" and $H_2 =$ "sickness" for $D =$ "young" and $\overline{D} =$ "older". The above assumption is saying that the events that belong to the same cause of death among different populations are more similar to each other than events that do not share the same cause of death nor the population (e.g. a young death by accident is more similar to an older death by accident than an older death by sickness). This assumption is not only satisfied in the set-up of Proposition 5, but also for general similarity functions that increase in the number of overlapping features.

Applying Proposition 2 yields the following:

$$\frac{\hat{\pi}(H_1|D)}{\hat{\pi}(H_2|D)}$$

$$= \frac{\pi(H_1D)}{\pi(H_2D)}\left[\frac{\pi(H_2D) + S(H_1D, H_2D)\pi(H_1D) + S(H_2D,\ H_2\overline{D})\pi(H_2\overline{D}) + S(H_2D,\ H_1\overline{D})\pi(H_1\overline{D})}{\pi(H_1D) + S(H_1D, H_2D)\pi(H_2D) + S(H_1D,\ H_1\overline{D})\pi(H_1\overline{D}) + S(H_1D,\ H_2\overline{D})\pi(H_2\overline{D})}\right]$$

Setting $\psi = \frac{\hat{\pi}(H_1|D)}{\hat{\pi}(H_2|D)} \Big/ \frac{\pi(H_1D)}{\pi(H_2D)}$ as the distortion in the likelihood ratio, note that individuals overestimate $\pi(H_1|D)$ if and only if $\psi > 1$. This occurs if and only if:

$$(S(H_2D,\ H_2\overline{D}) + S(H_1D,\ H_1\overline{D}) - S(H_2D,\ H_1\overline{D}) - S(H_1D,\ H_2\overline{D}))\pi(H_2|\overline{D})$$
$$+ \Big(1 + S(H_2D,\ H_1\overline{D}) - \big(S(H_1D, H_2D) + S(H_1D,\ H_1\overline{D})\big)\Big)$$
$$> 2 \cdot \big(1 - S(H_1D,\ H_2D)\big)\pi(H_1|D)$$

Now, by our assumption on similarity, note that both coefficients on $\pi(H_2|\overline{D})$ and $\pi(H_1|D)$, $(S(H_2D,\ H_2\overline{D}) + S(H_1D,\ H_1\overline{D}) - S(H_2D,\ H_1\overline{D}) - S(H_1D,\ H_2\overline{D}))$, and $\big(1 - S(H_1D,\ H_2D)\big)$, are positive. Consequently, this implies that, for a general similarity structure, one has overestimation of $\pi(H_1|D)$ ceteris paribus if:

a) $\pi(H_1|D)$ decreases below a certain threshold

b) $\pi(H_1|\overline{D})$ decreases below a certain threshold

For the similarity function $S(e_k, e_{k'}) = \delta^{\Sigma_i|f_{ki} - f'_{ki}|}$, the above inequality further simplifies to:

$$2\delta \cdot \big(1 - \pi(H_1|\overline{D})\big) + (1 - \delta) > 2\pi(H_1|D).$$

Rearranging gives Equation (9). ■

Corollary 3. The result follows from the definition of under/overreaction in Section 3.3 and inspection of Equation (9). ■

A2: Further results

Generalization to multiple hypotheses

In this section, we sketch out the extension of our model to the agent assessing multiple hypotheses. In this case, we assume $H_{i=1,\dots J}$ partition D, with $E = D \cup \overline{D}$. We assume now that the agent goes through the train of thought for each of the $J > 2$ hypotheses. Then, the ease of retrieval for hypothesis j is given by:

$$r(H_i) = \frac{\pi(H_i)}{\pi(H_i) + \sum_{\{j \neq i\}} \frac{S(H_i, H_j)}{S(H_i, H_i)}\ \pi(H_j) + \frac{S(H_i, \overline{D})}{S(H_i, H_i)}\ \frac{\pi(\overline{D})}{\pi(D)}}.$$

Then, as before, the agent's assessment of hypothesis is given by

$$\hat{\pi}(H_j) = \frac{R_j}{\sum_k R_k}.$$

Finally, regarding the generalization of Proposition 2 for $J > 2$ hypotheses, the result regarding the mean follows from an argument regarding law of large numbers. Again, a small detail is to assume that $T \mapsto \infty$ such that the probability of $\sum_k R_k = 0$ becomes vanishingly small, holding J fixed. Then, note that we have:

$$\hat{\pi}(H_j) = \frac{R_j}{R_j + \sum_{k \neq j} R_k} \mapsto_p \frac{r(H_j) + z_j/\sqrt{T}}{r(H_j) + \sum_{k \neq j} r(H_k) + z_{-j}/\sqrt{T}},$$

Where $z_j \sim N(0,\ r(H_j)\left(1 - r(H_j)\right),\ z_{-j} \sim N\left(0,\ \sum_{k \neq j} r(H_k)(1 - r(H_k))\right)$.

The rest of the derivation proceeds similarly.

Equivalence between under and overreaction and CG

In this section, we prove that the CG notion of under/overreaction, which is given by the positive/negative co-movement of forecast revision with forecast errors, is equivalent to our notion of under and overreaction in our setting.

Recall that our notion of overreaction/underreaction is when the DM overestimates H_1, $\hat{\pi}(H_1|D) > \pi(H_1|D)$, if the data D is objectively informative of H_1, $\pi(H_1|D) > \pi(H_1)$. In our setting, the CG notion of under/overreaction translates to the following: the DM overreacts if conditional on the DM *revising* the probability of H_1 up in response to $(\hat{\pi}(H_1|D) > \hat{\pi}(H_1))$, he overshoots, with a negative forecast error: $\hat{\pi}(H_1|D) > \pi(H_1|D)$.[37]

Consequently, to show that these two notions are equivalent, it suffices to prove that the DM revises his beliefs in the same direction as given by the update in the objective probabilities:

$$\hat{\pi}(H_1|D) > \hat{\pi}(H_1) \Leftrightarrow \pi(H_1|D) > \pi\left(H_1\middle|\overline{D}\right).$$

Using Equation (5), one can show that the above is equivalent to:

$$\frac{r(H_1 \cap D)}{r(H_2 \cap D)} > \frac{r(H_1)}{r(H_2)} \Leftrightarrow \pi(H_1|D) > \pi\left(H_1\middle|\overline{D}\right).$$

Denote $D_1 = D$ and $D_2 = \overline{D}$. We can then express the similarities involved in computing the probability estimates as:

$$S\left(H_i \cap D_l, H_j \cap D_m\right) = \frac{1}{|H_i \cap D_l||H_i \cap D_m|}\sum\nolimits_{u \in H_i \cap D_l v \in H_j \cap D_m} S(u,v)$$

$$= \delta^{|i-j|+|l-m|} \qquad for\ i,j,l,m = 1,2$$

$$S(H_i, H_i) = \frac{\pi(H_i|D_i)^2\pi(D_i)^2 + 2\delta\pi(H_i|D_i)\pi\left(H_i|D_j\right)\pi(D_i)\pi\left(D_j\right) + \pi\left(H_i|D_j\right)^2\pi\left(D_j\right)^2}{\pi(H_i)^2},$$

[37] The case when $\hat{\pi}(H_1|D) < \hat{\pi}(H_1)$ proceeds analogously.

$S(H_i, H_j)$

$$= \delta \frac{\pi(H_i|D_i)\pi(H_j|D_i)\pi(D_i)^2 + \delta\pi(D_i)\pi(D_j)\left[\pi(H_i|D_j)\pi(H_j|D_i) + \pi(H_i|D_i)\pi(H_j|D_j)\right] + \pi(H_i|D_j)\pi(H_j|D_j)\pi(D_j)^2}{\pi(H_i)\pi(H_j)}.$$

Define $\pi \equiv \pi(H_1|D),\ \overline{\pi} \equiv \pi\left(H_1\middle|\overline{D}\right),\ \varphi \equiv \frac{\pi(\overline{D})}{\pi(D)}$. It is then possible to write:

$$r(H_1 \cap D) = \frac{\pi}{\pi(1-\delta) + \overline{\pi}\varphi\delta(1-\delta) + \delta + \varphi\delta^2},$$

$$r(H_2 \cap D) = \frac{1-\pi}{(1-\pi)(1-\delta) + (1-\overline{\pi})\varphi\delta(1-\delta) + \delta + \varphi\delta^2},$$

$$r(H_1) = \frac{\pi^2 + 2\pi\overline{\pi}\delta\varphi + \overline{\pi}^2\varphi^2}{\pi[\pi + \delta(1-\pi)] + \delta\varphi[\delta(\overline{\pi}+\pi) + 2\overline{\pi}\pi(1-\delta)] + \overline{\pi}[\overline{\pi} + \delta(1-\overline{\pi})]\varphi^2},$$

$r(H_2)$

$$= \frac{(1-\pi)^2 + 2\delta(1-\pi)(1-\overline{\pi})\varphi + (1-\overline{\pi})^2\varphi^2}{(1-\pi)[(1-\pi) + \delta\pi] + \delta\varphi[\delta(2-\overline{\pi}-\pi) + 2(1-\overline{\pi})(1-\pi)(1-\delta)] + (1-\overline{\pi})[(1-\overline{\pi}) + \delta\overline{\pi}]\varphi^2}.$$

Remember that it suffices to prove the following claim:

$$\frac{r(H_1 \cap D)}{r(H_2 \cap D)} > \frac{r(H_1)}{r(H_2)} \Leftrightarrow \pi > \overline{\pi}$$

From the above expressions, one can see that there exist functions $Z(.,.), \Delta(.,.)$ such that:

$$r(H_1 \cap D) = Z(\pi, \overline{\pi}),$$

$$r(H_2 \cap D) = Z(1-\pi, 1-\overline{\pi}),$$

$$r(H_1) = \Delta(\pi, \overline{\pi}),$$

$$r(H_2) = \Delta(1-\pi, 1-\overline{\pi}),$$

so that the original inequality can be written as:

$$\frac{Z(\pi, \overline{\pi})}{Z(1-\pi, 1-\overline{\pi})} > \frac{\Delta(\pi, \overline{\pi})}{\Delta(1-\pi, 1-\overline{\pi})}.$$

We can write the above inequality as:

$$\frac{Z(\pi,\overline{\pi})}{\Delta(\pi,\overline{\pi})} > \frac{Z(1-\pi,1-\overline{\pi})}{\Delta(1-\pi,1-\overline{\pi})}.$$

First, it is immediate to evaluate:

$$\frac{Z(\overline{\pi},\overline{\pi})}{\Delta(\overline{\pi},\overline{\pi})} = \frac{Z(1-\overline{\pi},1-\overline{\pi})}{\Delta(1-\overline{\pi},1-\overline{\pi})} = \frac{1}{1+\delta\varphi}.$$

It suffices to show now that for $\pi > \overline{\pi}$, we have that:

$$\frac{Z(\pi,\overline{\pi})}{\Delta(\pi,\overline{\pi})} > \frac{1}{1+\delta\varphi} > \frac{Z(1-\pi,1-\overline{\pi})}{\Delta(1-\pi,1-\overline{\pi})}.$$

Consider first the left-hand side inequality. This is equivalent to requiring that, for $\pi > \overline{\pi}$:

$$Z(\pi,\overline{\pi})(1+\delta\varphi) - \Delta(\pi,\overline{\pi}) > 0.$$

Expanding the terms, it is possible to find that:

$$Z(\pi,\overline{\pi})(1+\delta\varphi) - \Delta(\pi,\overline{\pi}) = (\pi-\overline{\pi})K(\pi,\overline{\pi}),$$

where $K(\pi,\overline{\pi})$ is a second order polynomial with positive coefficients. Because $K(\pi,\overline{\pi}) > 0$ for any $(\pi,\overline{\pi}) \neq (0,0)$, it then follows that $(\pi-\overline{\pi})K(\pi,\overline{\pi}) > 0$ if and only if $\pi > \overline{\pi}$, which proves the claim. Similarly, the right hand side inequality reduces to proving that for $\pi > \overline{\pi}$:

$$Z(1-\pi,1-\overline{\pi})(1+\delta\varphi) - \Delta(1-\pi,1-\overline{\pi}) < 0.$$

Again, expanding the terms, it is possible to find that:

$$Z(1-\pi,1-\overline{\pi})(1+\delta\varphi) - \Delta(1-\pi,1-\overline{\pi}) = -(\pi-\overline{\pi})K(\pi,\overline{\pi}),$$

which, together with our previous argument, proves the claim.

Appendix B: Noise

The model yields novel predictions about "noise", that is, variability in beliefs given the same experiences and the same question (Kahneman et al. 2021). As shown in Proposition 2, noise naturally arises from sampling variation in recall. We now study noise in the experimental data and compare it with our model's predictions. The results are for the most part correlational but provide further, if suggestive, evidence that memory underlies the causal effects of Section 4.

We saw in Figure 4 that beliefs and recall are both noisy (i.e., heterogeneous across individuals) and positively correlated, as predicted. Our model makes further, and finer, predictions about this relationship, described in Proposition B.1.

Proposition B.1:

1. *The variance of* $\hat{\pi}(H_i)$ *decreases in* T*, the total recall attempts for each hypothesis.*
2. *If* $\hat{\pi}(H_i) \in \left[\sqrt{5}-2, 3-\sqrt{5}\right]$*, the variance of* $\hat{\pi}(H_i)$ *decreases in* $S(H_i, H_i)$ *and* $S(H_j, H_j)$.

First, the level of noise is negatively correlated with recall: as the number of recall attempts T gets larger, the share of successes for each hypothesis converges to its expected value, resulting in less heterogeneity. Second, if beliefs are not too far from 50:50, making hypotheses more self-similar should reduce noise in assessment. This is intuitive: when hypotheses are more self-similar, their recall is more successful, which increases sample size, in turn reducing the variability of beliefs.[38] Conversely, noise increases if hypotheses are less self-similar.

[38] The fact that average odds need to be sufficiently close to 50:50 highlights an important non-monotonicity in the relationship between heterogeneity and noise: if the heterogeneity of a particular hypothesis is sufficiently high, its likelihood will converge to zero, eventually reducing the level of noise. This is however not the case in our experiment, in which probability judgments are far from corners.

To assess the first prediction, Figure 9 shows the cross-sectional relationship between the total number of words recalled and the conditional variance of the fraction of recalled words that are animals (Panel A) and of beliefs (Panel B), pooling the four treatments of Experiment 1. We view the total number of words recalled by a subject as a proxy for T (so subjects are allowed to have different sample lengths). Consistent with Proposition B.1, we see a negative and statistically significant relationship in both cases ($p < 0.01$ for both). More sampling yields less variability in relative recall of hypotheses and, correspondingly, in beliefs.[39,40] In Appendix B, we show that another proxy for T—the time subjects spend answering the beliefs question—is also negatively correlated with the variance both of the fraction of recalled animals and of beliefs.

[39] To test the statistical significance for both dependent variables in Figure 9, we estimate by maximum likelihood a model in which the mean changes linearly and the conditional variance changes multiplicatively in total number of recalled words. The dashed curves in Figure 9 show predicted conditional variances from these estimates.

[40] The negative relationship shown in Panel A of Figure 9 could in principle be partly mechanical—if a subject correctly recalls all 30 words, it must be the case that 40% of them were animals. Note, however, that the recall task occurs *after* the beliefs question, so subjects cannot consult their list of recalled words when forming their probabilistic judgments. The negative relationship in Panel B is therefore not mechanical, and in any case we see a negative relationship in Panel A even for participants who recall only a small fraction of the words.

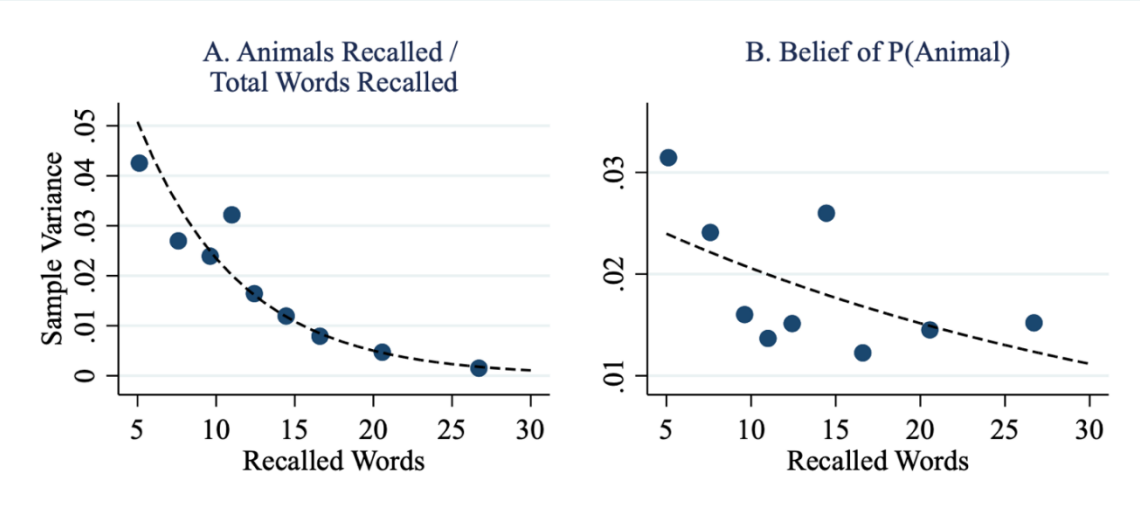

Figure 9: The Relationship between Recall and Noise

Notes: Panel A shows the sample variance of the fraction of recalled words that are animals (y-axis), conditional on decile of total number of recalled words (x-axis). Panel B shows the sample variance of beliefs about the probability animals, also conditional on decile of total number of recalled words (x-axis). The dotted lines show predicted values from maximum likelihood estimates of a model where the mean of the dependent variable varies linearly and the conditional variance multiplicatively in the total number of recalled words. Both panels restrict the data to *T1* and *T3*, where the true frequency of animals is 40%.

We can also exploit cross-treatment differences in similarity to test the second prediction in Proposition B.1. Recall that the hypothesis alternative to animals is much less self-similar in T3 than in T2 (where non-animals are all people's names). Proposition B.1 holds that decreasing retrieval fluency for non-animals should increase variance in the relative recall of animals (i.e., recall of animals divided by recall of all words) and, therefore, raise the variance in beliefs. Figure 10 shows that, indeed, there is greater variance in the fraction of recalled words that are animals in T3 than in T2 ($p < 0.01$) and correspondingly, greater variance in beliefs ($p < 0.01$).

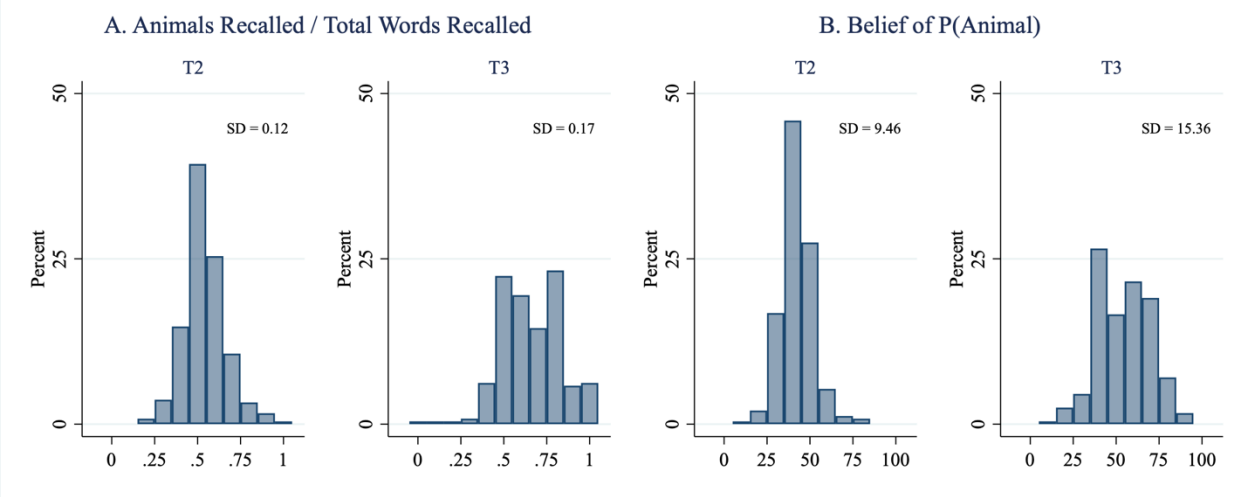

Figure 10: Treatment effects on Noise

Notes: Panel A shows the distribution of the fraction of recalled words that are animals in T2 (leftmost graph) and in T3 (second graph to the left). Panel B shows the distribution of beliefs about the probability of animals in the same treatments.

In sum, selective recall seems to be a promising avenue to think about systematic biases in probability judgments and belief heterogeneity.

Proof of Proposition B.1. We can work out the distribution of $\hat{\pi}(H_i)$ starting from the characterization of the odds in Proposition 2. Using the same logic of the proof of Proposition 2, for large enough T the odds of H_j relative to H_i are distributed as:

$$\frac{N_j}{N_i} \to N\left(\frac{r(H_j)}{r(H_i)}, \frac{1}{T}\cdot\left(\frac{r(H_j)^2}{r(H_i)^4}\cdot r(H_i)\big(1-r(H_i)\big) + \frac{1}{r(H_i)^2}\cdot r(H_j)\big(1-r(H_j)\big)\right)\right).$$

By applying the delta-rule on $f(\epsilon) = \frac{1}{1+x+\epsilon} = \frac{1}{1+x} - \frac{1}{(1+x)^2}\,\epsilon + O(\epsilon^2)$, where $x = \frac{r(H_j)}{r(H_i)}$ is the mean above and ϵ is the Gaussian discrepancy from it, we find that $\hat{\pi}(H_i)$ follows the asymptotic distribution:

$$\hat{\pi}(H_i) = \frac{1}{1+\frac{N_j}{N_i}} \sim N\left(\frac{r(H_i)}{r(H_i)+r(H_j)}, \frac{1}{T}\cdot\frac{r(H_i)^2 r(H_j)^2}{\big[r(H_i)+r(H_j)\big]^4}\left[\frac{1-r(H_i)}{r(H_i)} + \frac{1-r(H_j)}{r(H_j)}\right]\right).$$

The model thus predicts noise in $\hat{\pi}_i = \hat{\pi}(H_i)$ to be equal to be a function of recall fluencies:

$$v(\hat{\pi}_i) = \frac{1}{T}\cdot\frac{r_i^2 r_j^2}{(r_i+r_j)^4}\left(\frac{1-r_i}{r_i} + \frac{1-r_j}{r_j}\right),$$

where $r_i = r(H_i)$. This can be written as:

$$\nu(\hat{\pi}_i) = \frac{1}{T} \cdot \frac{r_i r_j \left(r_i + r_j - 2 r_i r_j\right)}{\left(r_i + r_j\right)^4}.$$

Clearly the expression fulfils $\nu(\hat{\pi}_i) = \nu\left(\hat{\pi}_j\right) = \nu(1 - \hat{\pi}_i)$.

To prove the first property, it is evident that $\frac{\partial \nu(\hat{\pi}_i)}{\partial T} < 0$. For the second property, after some algebra one can find that:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_i} \propto r_j^2 (1 - 4 r_i) + r_i r_j (4 r_i - 3) - 4 r_i^2,$$

Define $R = r_i + r_j$. After some algebra one can find:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_i} \propto 1 - 4\hat{\pi}_i - \hat{\pi}_i^2 - 4\hat{\pi}_i (1 - \hat{\pi}_i)^2 R.$$

A sufficient condition for the above derivative to be negative is obtained by imposing $R \mapsto 0$, which yields:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_i} < 0 \quad if \quad \hat{\pi}_i > \sqrt{5} - 2,$$

which also yields:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_j} < 0 \quad if \quad \hat{\pi}_i < 3 - \sqrt{5}.$$

Appendix C: Details of Experimental Design

In this section, we describe the design of the two experiments in greater detail. Both experiments were pre-registered on the AEA RCT Registry, with ID AEARCTR-0006676.

Design of Experiment 1

Experiment 1 was conducted in March of 2021 among Bocconi undergraduates. We recruited participants via email from the experimental economics listserv. Participants completed the experiment online from home (due to Covid restrictions). They earned a 4 euro Amazon gift card as an incentive to take the survey, and had the chance to earn an addition 2 euro bonus for correct responses. The median respondent took less than 10 minutes to complete the survey. In total, 1,200 respondents participated in the survey (this sample size was preregistered), roughly 240 for each of the five treatments.

Participants could choose to take the survey in either Italian or English, of which 19% chose the latter. Both the text of the questions and the words in the memory task were translated according to participants' choice. After a consent form, participants were told that the survey would include several questions for which they could earn a 2 euro bonus for answering correctly. At the end of the experiment, the computer would randomly choose one of these questions as the "Bonus Question" to base their bonus on.

Participants were then told they would be shown a series words, one by one in a random order, which would take about a minute. They were told that the Bonus Question would be about these images, and were therefore encouraged to pay attention. They then answered three simple comprehension questions, with corrections for errors. After the words, participants had to wait 10

seconds before proceeding to the questions. They were encouraged to “Take a moment to reflect on the words you saw.”

They were then asked the probabilistic question about the percent chance that it fell into various categories, as described in the main text. They were told that, if this question was chosen as the Bonus Question, they would earn the 2 euro bonus if their answer was within 5 percentage points of the right answer. Participants then saw a confidence question asking they how certain they were that their previous answers were within 5 percentage points of the right answer.

Participants then answered the free recall questions, which asked them to “please list up to 15 [category] that you remember seeing in the list of words we showed you. You do **not** have to fill in all 15 lines.” If a recall question was chosen as the Bonus question, participants’ were told that their “chance of earning the 2 euro bonus will increase by 10 percentage points for every correct answer, and it will decrease by 10 percentage point for every incorrect answer you’re your chance cannot go below zero or above 100%). Don’t worry about typos or misspellings.”

Participants were then asked a series of follow-up questions. The first asked about whether they felt they paid more attention to the early, middle, or late part of the sequence of words (the large majority say the early part). The next asked about whether they noticed that the words fell into any categories and prompted them to write what categories they remembered seeing.

Next, they were asked if, when watching the words, they wrote any of them down or took a video to refer to later. 13% report doing so, though excluding them from the analysis does not change any qualitative results (though their overall performance on the recall task is, unsurprisingly, better than average).

Finally, the survey asked about native language, age, gender, and whether they found the survey questions difficult or easy to answer and whether they found the instructions confusing or clear. The majority found it difficult (by far the most common reason given in a free response is that it was difficult to remember all the words) but 99% found the instructions very or moderately clear.

T5 treatment in Experiment 1

In addition to the treatments described in the main text, in Experiment 1 we also ran a treatment called T5. T5 was identical to the T2 except we replaced women's names in T2 with "ocean animals" (e.g., "Shark", "Whale", etc.). Note that, in T2, all the animals were "land animals" (e.g., "Lion", "Dog", etc). In addition, instead of asking about the probability of the randomly chosen word being an "animal" as in T2, in T5 the question asked for the probability that it was a "land animal." The free recall task in T5 correspondingly asked respondents to recall up to 15 "land animals." The purpose of this treatment was to increase the cross-similarity between the hypotheses under consideration. In T5, "Land Animals" and "Other" are arguably more similar than "Animals" and "Other" in T2. According to our theory, this should reduce the ease of recall of land animals in T5 compared to T2. Panel B of Figure C1 shows that, indeed, respondents correctly recall fewer land animals in T5 than in T2.[41]

[41] Whether there should be greater ease of recall for the other category in T2 than in T5 depends on whether men's names and women's names are more self-similar than men's names and ocean animals. In fact we find somewhat higher recall of non-land-animals in T5 than in T2 ($p = 0.07$), suggesting that if anything the opposite is true. An alternative explanation, outside the model, is that ocean animals are simply more memorable or salient than women's names. Consistent with the effects on recall, beliefs about the probability of land animals are slightly higher in T5 than in T2, though the difference is not statistically significant (point estimate = 1.8pp, $p = 0.20$).

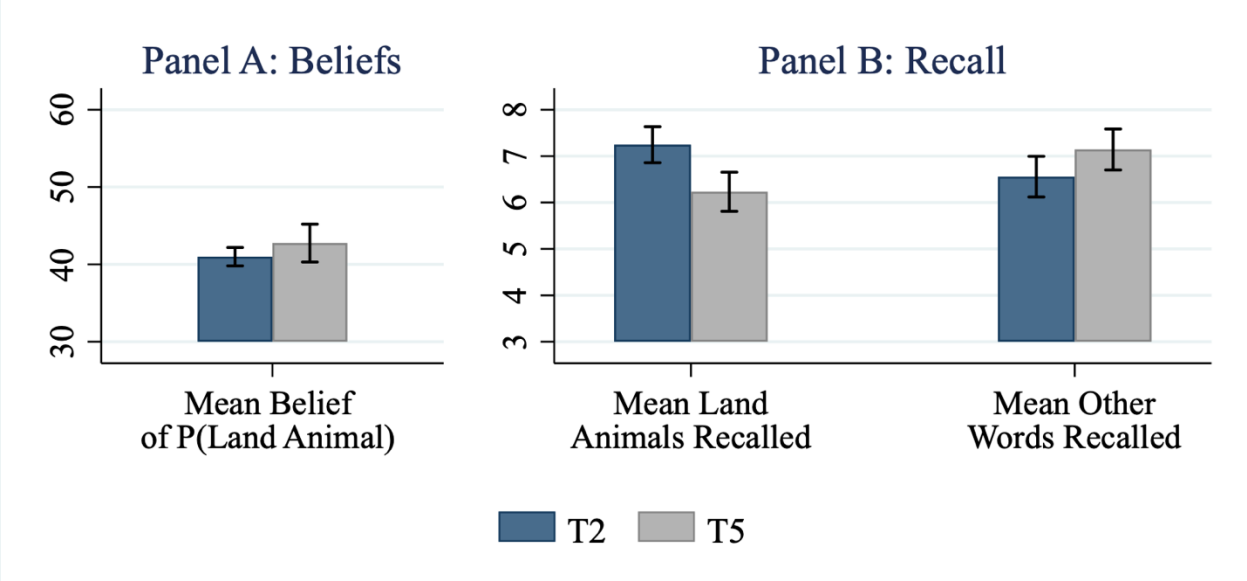

Figure C1: Beliefs and Recall in T2 vs T5

Notes: Panel A shows the average belief of the probability of (land) animals minus the true proportion in T2 and T5. Panel B shows the number of land animals and non-land-animals recalled in the free recall task in these treatments. Bands show 95% confidence intervals.

However, there appears to have been significant confusion about what qualified as an "ocean animal." 27% of respondents list an ocean animal when asked to recall a land animal (in contrast, in T2, no respondents list a name when asked to recall animals), suggesting that this distinction was less natural than we anticipated.

Additional Analyses of Experiment 1

Time as a Proxy for *T*

Figure C2 plots the relationship between the time subjects spent on the beliefs question (winsorized at 100 seconds) and the sample variance of the fraction of recalled words that were animals (Panel A) and of beliefs about P(Animal) (Panel B). We see a statistically significant ($p < 0.01$) negative relationship in both cases.

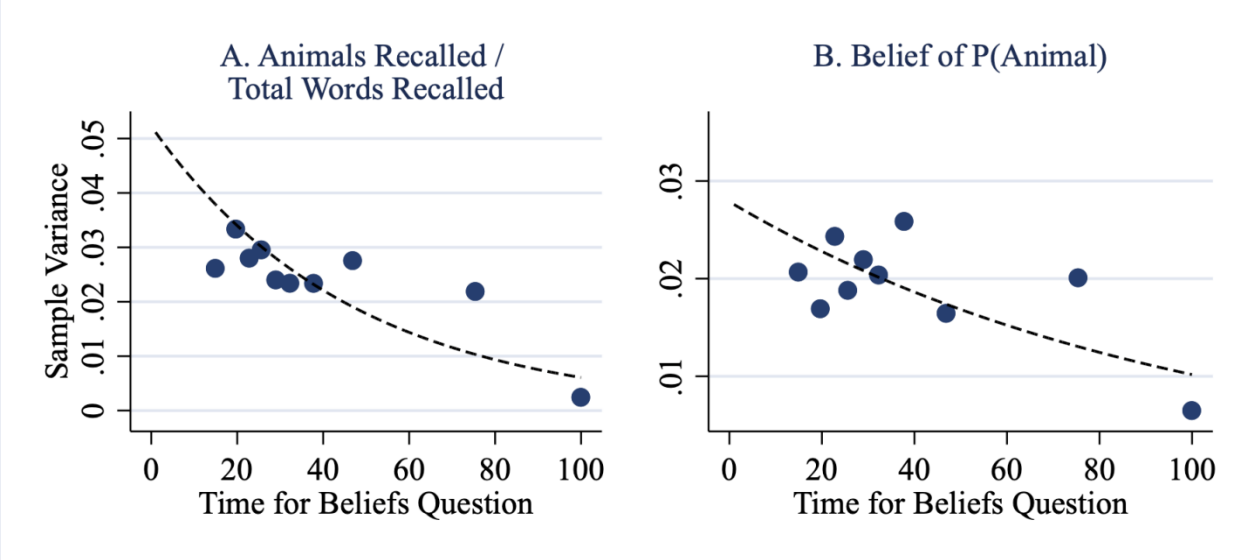

Figure C2: The Relationship between Recall Time and Noise

Notes: Panel A shows the sample variance of the fraction of recalled words that are animals (y-axis), conditional on decile of time spent on the beliefs question (x-axis). Panel B shows the sample variance of beliefs about the probability animals, also conditional on decile of time spent on the beliefs question (x-axis). The dotted lines show predicted values from maximum likelihood estimates of a model where the mean of the dependent variable varies linearly and the conditional variance multiplicatively in time spent on the beliefs question. Both panels restrict the data to *T2* and *T3*, where the true frequency of animals is 40%.

Recency Effects

Figure C3 plots the probability of recalling an exemplar as a function of its chronological position in the sequence.

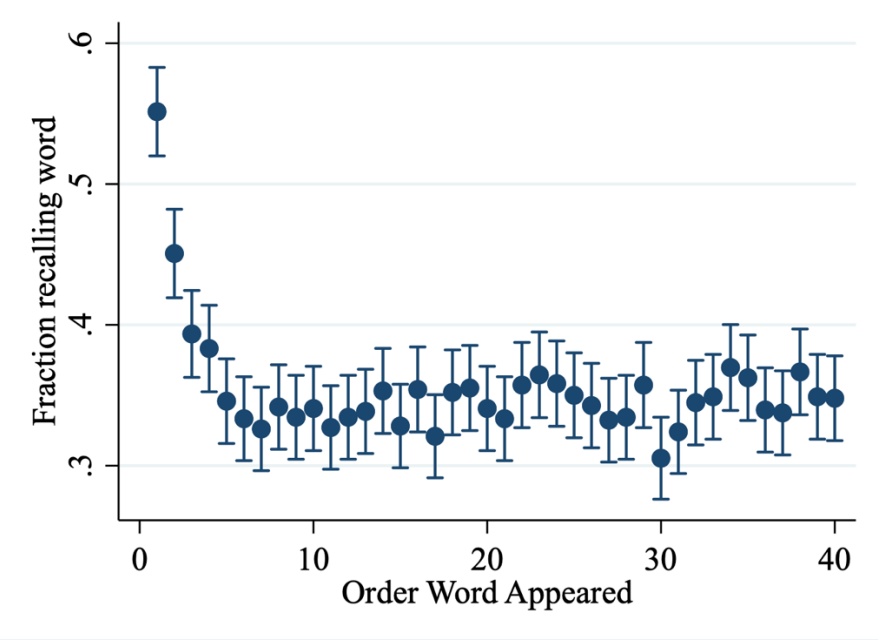

Figure C3: Primacy effect in Recall

Notes: Figure shows the proportion of respondents who listed a word in one of the free recall tasks (y-axis) depending on the chronological order in which the word appeared to them (x-axis). The order of words was randomized. Bands show 95% confidence intervals.

Consistent with existing work on memory (Kahana, 2012), there is a strong primacy effect in recall, in which early experiences are likelier to be recalled than late ones. If retrieval from memory reflects probability beliefs, then subjects who see many Animals early should have higher subjective probability of Animals, and vice versa. Contrary to the literature's finding, however, we do not find a strong recency effect, or the tendency to recall the latest entries. Indeed, regressing beliefs on the number of animals shown in the first 5 words shows that indeed those who happened to be shown more animals first, controlling for treatment, report a higher probability for animals:

$$P^{belief}(Animal) = Constant + .90\ (.28) \cdot \#\ Animals\ in\ First\ 5\ Words$$

We take this to be further evidence of the role of recall in shaping beliefs.[42]

[42] We take these results to be somewhat suggestive. For example, we do not find the analogous impact on recall: that is, people who see many animals at the beginning do not disproportionately remember animals. Second, our effect is concentrated in participants who get 4 or 5 animals in the first 5 words.

One might worry that if, by chance, participants in some treatments tended to see more animals early in the sequence, that our treatment effects might partly reflect primacy effects. Table C1 below regresses participants beliefs (columns 1-4) and the fraction of recalled words that were animals (columns 5-8) on treatment dummies, excluding T2 to keep the true number of animals shown constant. Columns 1 and 5 show that, without controlling for primacy effects, respondents have higher beliefs and recall more animals in T3, and that they have lower beliefs and recall fewer animals in T4. These are the effects that we discuss in the main text. Columns 2, 3, and 4 add dummies for each possible number of animals that appeared in the first one, five, and ten words that respondents saw. We see no meaningful effect on the average belief across treatments. Columns 6, 7, and 8 add similar dummies to the regression investigating recall, and we again see no change in the estimated coefficients. We conclude that our results are not substantially changed by the primacy effects that we uncover.

	Dep. Variable: Belief				Dep. Variable: Fraction Recalled Animals			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
T3	11.69***	11.69***	11.79***	11.77***	0.12***	0.12***	0.12***	0.12***
	(1.16)	(1.16)	(1.17)	(1.16)	(0.01)	(0.01)	(0.01)	(0.01)
T4	-2.18**	-2.18**	-2.23**	-2.04**	-0.03**	-0.03**	-0.03**	-0.03**
	(0.88)	(0.88)	(0.87)	(0.87)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	40.99***	40.99***	40.98***	40.89***	0.54***	0.54***	0.54***	0.54***
	(0.61)	(0.61)	(0.60)	(0.61)	(0.01)	(0.01)	(0.01)	(0.01)
Includes Dummies for Animals Among First…		Word	5 Words	10 Words		Word	5 Words	10 Words
N	719	719	719	717	718	718	718	716

Table C1: Controlling for Primacy Effects

Notes: Table shows OLS regressions. The dependent variable in columns 1 to 4 is participants' belief about the fraction of words that are animals, and in columns 5 to 8 is the fraction of words participants recalled that were animals. The data are restricted to T1, T3, and T4. Columns 2, 3, and 4 include a dummy variables for each possible number of animals among the first one, five, and ten words that the participant saw, repectively. Columns 6, 7, and 8 include similar dummies in the same order. Robust standard errors in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

At the end of the survey, we also asked participants "While seeing the words, did you feel that you paid more attention to the beginning, middle, or end of the sequence?" Table C2 below shows the fraction of respondents who gave each answer across our four treatments. While the majority of respondents report paying more attention to the beginning of the sequence (in line with the primacy effects on recall shown above), none of the differences across treatments are statistically significant.

	Beginning	Middle	End
T1	0.61	0.25	0.14
T2	0.64	0.19	0.16
T3	0.64	0.19	0.17
T4	0.61	0.21	0.17

Table C2: Attention Across Treatments

Notes: Table shows, across treatments, the fraction of respondents who reported paying more attention to the words at the beginning, middle, or end of the sequence.

Relationship to cognitive uncertainty and beliefs

By linking probability beliefs to measurable recall, our model also speaks to research connecting cognitive imprecision and uncertainty to probability judgments. Enke and Graeber (2019) find greater attenuation to 50-50 for agents who report a higher level of subjective, or

cognitive uncertainty. The authors show that their results can arise from a Bayesian processing of noisy cognitive signals – the greater the noise, the greater the subjective uncertainty and shrinkage towards the prior mean. However, beyond its indirect impact on subjective uncertainty and probability judgments, measuring cognitive noise remains an important challenge.

Our measurement of free recall in conjunction with probability elicitation provides a complementary insight into the underlying determinants of cognitive uncertainty. If people's probabilistic beliefs are informed by sampling from memory, an important source of noise and uncertainty can come from the inability to recall relevant data. To evaluate this approach, after participants give their probabilistic beliefs (but before the recall tasks), we ask them how confident they are that their answers are within five percentage points of the true answer (the threshold for earning a bonus for accuracy). They respond by dragging a slider that ranged from "Very Uncertain" to "Very Certain," which we convert to a 0-100 scale.

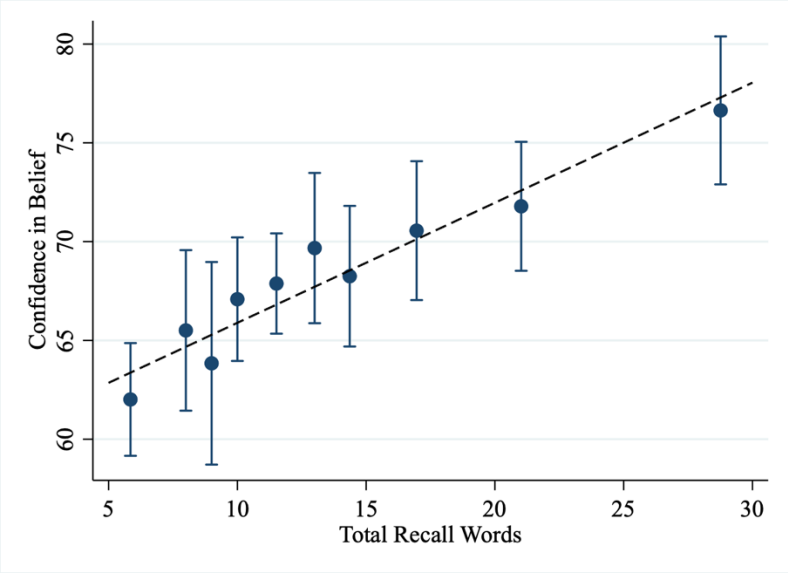

Figure C4: Confidence and Recall

Notes: This figure shows the average confidence in respondents' belief of the probability of animals, conditional on the total number of recalled words in the free recall tasks (binned by decile). Bands show 95% confidence intervals. The dashed line shows the OLS line of best fit.

Figure C4 shows a strong positive relationship between the total number of recalled entries and this measure of participants' confidence. The red line shows the OLS estimate, which is highly significantly positive ($p < 0.001$). To summarize, our exploratory analysis suggests a strong link between recall and subjective uncertainty. Our suggestive findings are consistent with the insight that failure to retrieve relevant information leads to higher subjective uncertainty: the limitation of memory is an important driver of subjective uncertainty.

Design of Experiment 2

Experiment 2 was conducted in November 2020, also online with Bocconi undergraduates recruited in the same was as Experiment 1. Participants earned a 4 euro Amazon gift card for completing the survey, plus a possible 2 euro bonus for accuracy. The median respondent took about 9 minutes to take the survey. In total 1,203 respondents participated in the experiment, approximately 150 per treatment. The pre-registered sample size was 1,200, though three additional respondents completed the survey before we closed it.

The survey had the same design as Experiment 1, with a few exceptions. First, there were no comprehension questions following the instructions. Second, respondents were told that there would be 40 images, each of which would be either a word or a number and would be blue or orange. After participants saw the words, there was not an enforced 10 second delay as in Experiment 1 (in practice the median respondent spent 4 seconds on the transition page after the images ended and before the questions about them).

After this, respondents were asked "Suppose the computer randomly chose an image from the images you just saw. It is orange. What is the percent chance it is a word? Please indicate your answer by clicking on the scale below and then dragging the slider." A slider below this text went from 0 to 100 (with no default starting value). The confidence question after this was the same as in Experiment 1.

After this question (which is the primary question we focus on), respondents were also asked an analogous question but assuming the randomly chosen image was blue. Finally they were asked the same question but supposing they did not know the color. Both of these questions also included a confidence question asking how confident they were that their answer was within 5 percentage points of the right answer.

Before the free recall question, respondents answered 4 questions of the following type: "Was X [this word appeared in either blue or orange] (in blue or orange) among the images you saw?" We chose 4 random words from the same list of time-related words but that were *not* among the images the respondent had seen. Two of these words appeared in blue and two appeared in orange, in a random order. If any of these questions was chosen to be paid on, the respondents earned the 2 euro bonus if they answered "no" (since this was always the correct answer).

The free recall question asked respondents to list up to 10 orange words that they remembered seeing. They were told that, if this question was chosen for payment, their chance of receiving the 2 euro bonus would increase by 10 percentage points for every correct answer and decrease by 10 percentage points for every incorrect answer (though it could not, of course, go below zero).

Finally, the survey ended with some questions including age, native language, how difficult they thought answering question in the survey was, whether they found the instructions confusing,

and whether they experienced any technical issues. The survey also presented a multiple choice questions asking respondents what, when they were viewing the images, they expected us to ask of them. 58% responded that they expected to be asked to list specific words, 39% expected to be asked to say how many images of different types (words, etc.) they saw, and only 17% expected to be asked about the colors of words. 38% said they "did not have anything in particular in mind".

Additional Analyses of Experiment 2

In Experiment 2, all respondents are first asked the probability that a randomly drawn image was a word conditional on it being orange. This question is the primary object of interest for our analysis. However, after this question, we also ask the probability that a randomly drawn blue image is a word and then the probability that a randomly drawn word unconditional on color is a word. For completeness, we report the average beliefs for these questions here. We did not ask corresponding recall questions for either blue words or for words unconditional on color.

Figure C5 shows the mean beliefs about the probability of words both conditional on blue (Panel A) and unconditional on color (Panel B). Panel A shows that while mean beliefs about the probability of words conditional on blue change dramatically between the H and L treatments, they do not go all the way to 100% and 0%, respectively. This may be because respondents inferred that we would not ask the question if the answer were trivial. When orange images are 50% words in the NM treatment (meaning that blue images are also 50% words), the average belief of P(Word | Blue) is 46.3 percent, lower than the truth ($p = 0.01$). For all other questions, the average belief is attenuated toward 50%, consistent with our theory's prediction that rare hypotheses should be exaggerated.

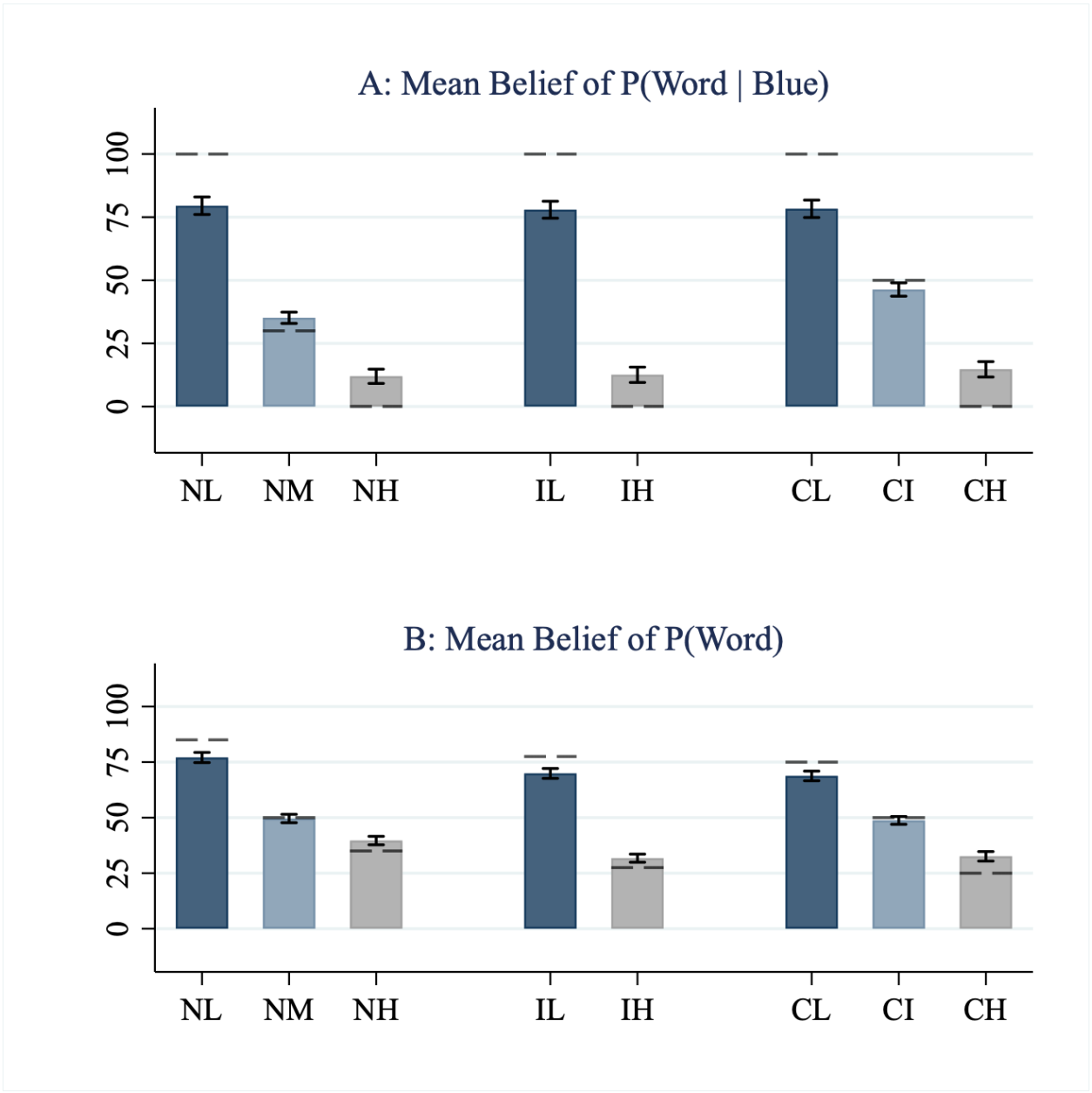

Figure C5: Other Beliefs in Experiment 2

Notes: Panel A shows the average belief of the probability of the randomly drawn image being a word conditional on it being blue. Panel B shows the average belief of the probability of the randomly drawn image being a word unconditional on its color. In the *L* treatments, all blue images were words. In the *H* Treatments, all blue images were numbers. In the *M* treatment when 70% of orange images are words (CM), 30% of blue images are words. In the *M* treatment when 50% of orange images are words (NM), 50% of blue images are also words. Bands show 95% confidence intervals. Dashed lines show the correct answer for each treatment.

In addition to the recall task described in the main text, in which respondents were asked to list orange words that they recalled seeing, the survey also included a "misrecognition" task. In particular, respondents were asked four questions, each asking them whether a specific word in a specific color was among the images they saw. Two of the named words were blue, and two were orange, and they appeared in a random order. In each case, the named word was *not* among the images they were previously shown.

Figure C6 below shows the average number of blue and orange words that respondents erroneously reported recognizing from the images they were shown. Overall, we do not see large or systematic differences in the number of orange words mis-recognized across treatments. We do see large differences across treatment in the number of blue words mis-recognized. This is perhaps less surprising, as the frequency of blue words changed much more dramatically across treatments than the frequency of orange words.

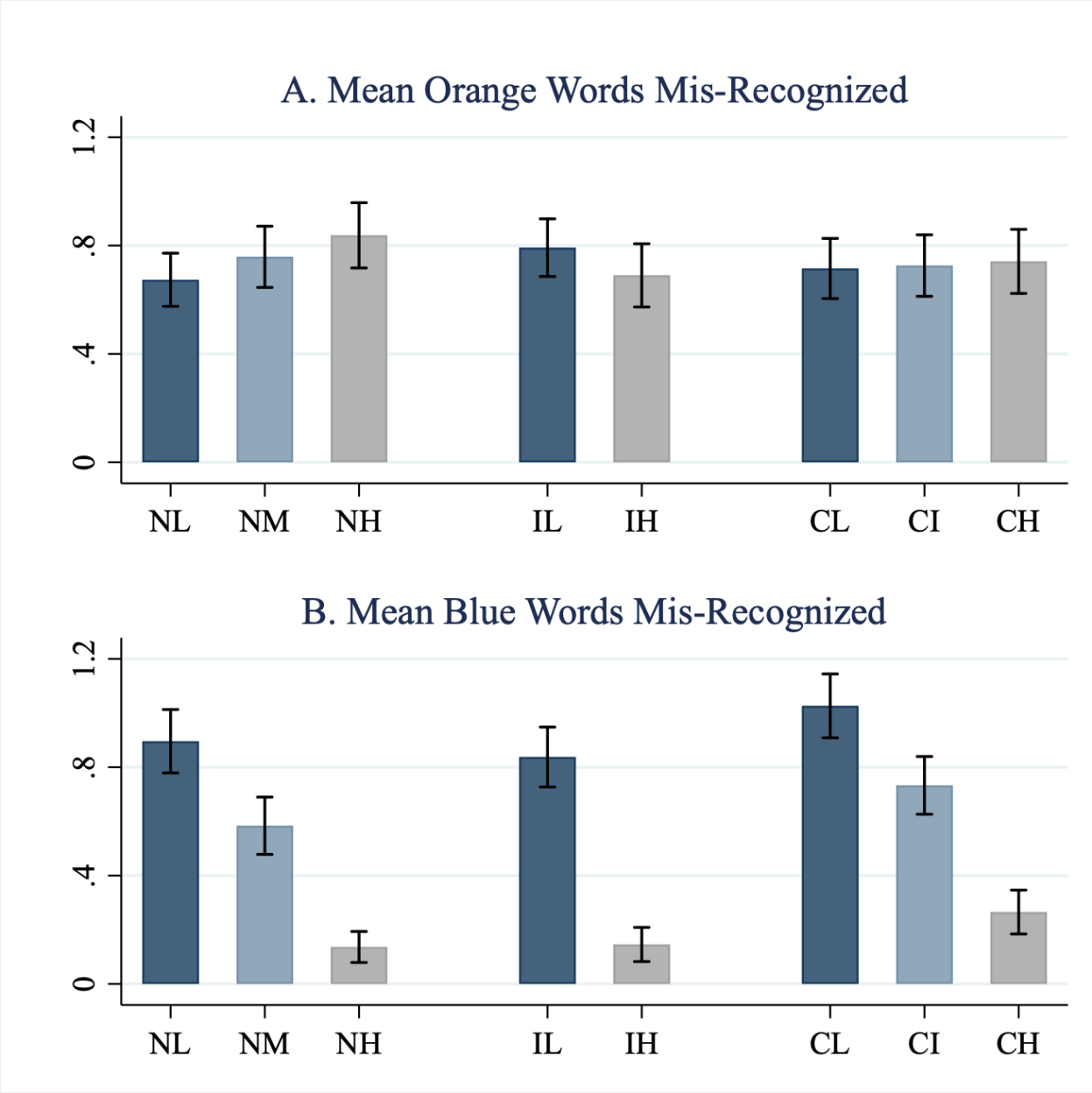

Figure C6: Mis-Recognition in Experiment 2

Notes: Panels A and B respectively show the average number of blue and orange words that respondents erroneously reported recognizing from the images they were shown. Bands show 95% confidence intervals.